ENVIEVAL

Development and application of new methodological frameworks for the evaluation of environmental impacts of rural development programmes in the EU

(Project Reference: 312071)

Area 2.1.4: Socioeconomic research and support to policies

KBBE.2012.1.4-08:

Report D7.2

Report on the cost-effectiveness of the evaluation approaches

---

Authors: **Anne Wolff (TI), Gerald Schwarz (TI), Bernhard Osterburg (TI), Frank Offermann (TI)**

- with input from the AUA, BEF, CREA, JHI, LUKE, SZIE teams

Approved by Work Package Manager of WP7: **Bernhard Osterburg (TI)**

Approved by Project Coordinator: **Gerald Schwarz (TI)**

Date: January 2016

This document was produced under the terms and conditions of Grant Agreement No. 312071 under the European Union's Seventh Framework Programme for research, technological development and demonstration. It does not necessary reflect the view of the European Union and in no way anticipates the Commission's future policy in this area.

Table of Contents

List of Figures

List of Tables

Executive summary

Two of the main aims of the cost-effectiveness assessment of the evaluation methods in WP7 are: a) to estimate the cost of the required resources for indicators and evaluation methods and to analyse the determinants of the costs; and b) to assess the effectiveness of the developed indicators and evaluation methods based on the case studies. The overall purpose of the assessment is:

- to provide guidance on the cost-effective application of the indicators and evaluation methods in future evaluations and
- to gain more insights into the impact of monitoring and data requirements and efforts on the cost-effectiveness of RDP evaluations.

In order to achieve these objectives a systematic and structured approach is required to identify and quantify all cost components of the development and application of the tested evaluation methods, and to develop a conceptual framework for the qualitative assessment of the effectiveness of the tested indicators and evaluation methods, applied in the public good case studies in WP6. While the bulk of the required information has been directly collated through the project team, stakeholder interviews and workshops (as well as sessions at the annual SRG meetings) with representatives from monitoring organisations and evaluators were required to collate information on monitoring programmes and to validate cost and effectiveness assessments of the project team.

The report outlines the conceptual framework developed for the cost-effectiveness assessment, summarises results of the cost and effectiveness assessment of the evaluation approaches tested in the case studies paying particular attention to the implications and role of different stakeholder priorities for the assessment of the effectiveness of evaluations, and provides synopsis of the implications of different data and monitoring programme scenarios on the cost-effectiveness of the tested evaluation approaches.

The ENVIEVAL project has tested a structured approach to assess the cost and performance of the evaluation approaches for different public goods in the case studies. The identified costs of the required resources were collected for each tested evaluation approach. The cost templates could help evaluators to plan and control evaluation cost in a structured way and to identify the main drivers of cost. The comparison of costs of evaluation approaches remains challenging although the detailed assessment of cost helps to show the drivers of cost for each evaluation approach.

Comparability is further limited due to different conditions in the partner countries (e.g. different data access and expertise for statistical analysis) and evaluation agencies. This shows that the mere comparison of cost of evaluation approaches is not sufficient. But what is important is to raise the awareness of what suitable and advanced evaluation including adequate environmental programmes cost. This has also been particularly highlighted in the stakeholder workshops. It is also important to consider the effectiveness of the approaches in order to get a holistic valuation of the cost-effectiveness of evaluation approaches.

The summary of the performance assessment of the tested evaluation approaches highlighted how the different stakeholder priorities affect the interpretation of the results and ultimately the selection of the approach for environmental impact evaluations of RDPs. The results of the effectiveness or performance assessment can be differently interpreted depending on the set of priorities attached to the judgement criteria and the final decision which evaluation approach to select often depends on the particular priorities of the stakeholders. The final selection revolves around an inspection of the performance assessment considering the relative costs of the different approaches as well as specific circumstances, preferences and abilities of the end-user (stakeholder). It is however important that a consistent framework is used with clearly defined criteria and performance or impact levels is used.

The identification of stakeholder priorities and their different weights for judgement or effectiveness criteria of evaluation approaches is important for an ex-ante assessment of the potential contributions of possible approaches to select for the evaluation of environmental of RDPs, informing the selection of evaluation approaches. The explicit consideration of different stakeholder priorities also contributes to a better understanding to what extent the applied evaluation approaches have delivered the required results, addressed existing evaluation challenges and helps to identify the need for further improvements in both the data infrastructure and methodological development. The development of the conceptual framework with a set of quality and judgement criteria as well as performance levels is the basis for a robust and sound assessment of the effectiveness of evaluation approaches. The framework developed in the ENVIEVAL project has attempted to fill the gap of a lacking framework and provides a starting point for further improvements of effectiveness assessments of environmental evaluations of RDPs.

Detailed assessments of the performance of the tested evaluation approaches using the framework developed in section 3.2 have been reported in the case study summary reports in Deliverable D6.3. Here only a short summary of the performance matrix of the tested evaluation approaches is

provided. The high performance levels for the Establishment of causal relationships and the Appropriateness of indicators and methods to capture the complexity of environmental relationships highlights the emphasis of the public good case studies on contributions of additional (non-CMES) indicators tested to address indicator gaps and contributions of advanced modelling approaches tested at micro and macro level for dealing with the complexity of public goods (see also the discussion section of Deliverable D6.3). In contrast, only 3, respectively 4, tested evaluation approaches achieved a high performance level for the criteria Establishment of consistent micro-macro linkages and Assessment of net-impacts, which reflects the severity of the methodological challenges underlying those two criteria as well as the large data requirements of evaluation approaches able to address these challenges.

During the evaluation process different decisions along the steps of the logic model influence the cost and effectiveness of the evaluation approaches. It can be concluded that in all evaluation steps decisions have to be made that influence the cost-effectiveness of the evaluation approaches. Particularly decisions in the beginning of the evaluation process and related to data availability have impacts on the overall effectiveness of the evaluation as they influence data generation, database development and the application of the evaluation method. However, good decisions in the beginning cannot provide good evaluation results if later decisions in the evaluation process inhibit the analysis. Thus, a balanced and considerable resource use could help to facilitate a successful evaluation. This is hampered by the limited funding and time restrictions that are available for evaluation. A realistic cost calculation and a targeted evaluation could help to overcome these restrictions.

The implementation of the monitoring cost scenarios for selected case studies of the ENVIEVAL project show that an improvement of the effectiveness of evaluation approaches can be achieved with relatively low cost, at least if one puts the additional cost into the context of the overall RDP budget. Also, small efforts such as the integration of alternative existing data sets or a more detailed analysis and processing of available data can already improve the effectiveness of evaluations. Further cost savings can be achieved by embedding additional data collection, or more generally, environmental monitoring for the evaluations of RDPs into a multi-purpose monitoring system.

If additional data collection is necessary to improve the evaluation method, cost are usually higher as data collection is costly and requires more efforts. The improvements either enable the use of advanced counterfactual methods or increase the cost-effectiveness of using those methods.

Advanced counterfactual methods are crucial to be able to assess net impacts of RD measures. Improved monitoring data is also needed for the assessment of synergies between measures to enable the analysis of multiple comparison groups. Further, the improvements meet largely the stakeholder priorities identified in national stakeholder workshop in the partner countries. This is a validation that the cost scenarios address the main evaluation challenges of the particular case study setting.

Whether the developed scenarios and their results are transferable to other cases requires further validation. The transferability for indicators that are applied across member states (e.g. the farmland bird index) is probably higher than for country-specific situations. However, the improvements of the different scenarios show ways of enhancing data quality and/or quality which are expected to be useful for monitoring data for varying indicators or methods. A number of lessons can be derived for future environmental monitoring programmes:

- Setting data pre-requisites at the beginning of each programming period facilitates sound statistical analyses of environmental impacts and robust recommendations
- Planning of impact evaluations at the stage of scheme design helps to ensure necessary data availability for consistent evaluation
- Adjustments to sampling and monitoring methods targeted at RDP evaluation can improve cost-effectiveness of the evaluation process
- Embedding additional data collections for improving RDP evaluations into a multi-purpose monitoring system eventually leads to recourse savings of the public sector and more comprehensive data sets.

To improve the cost-effectiveness of environmental evaluations it would be useful to consider data requirements and evaluation needs from the beginning of the evaluation process. Thus, it is recommended to develop the monitoring system jointly with the RD measures in order to be able to more reliably prove the environmental impacts of the measures. This would facilitate the application of statistic-based evaluation methods and increase the cost-effectiveness of environmental evaluations of rural development programmes.

1 Introduction

Two of the main aims of the cost-effectiveness assessment of the evaluation methods in WP7 are: a) to estimate the cost of the required resources for indicators and evaluation methods and to analyse the determinants of the costs; and b) to assess the effectiveness of the developed indicators and evaluation methods based on the case studies. The overall purpose of the assessment is:

- to provide guidance on the cost-effective application of the indicators and evaluation methods in future evaluations and
- to gain more insights into the impact of monitoring and data requirements and efforts on the cost-effectiveness of RDP evaluations.

In order to achieve these objectives a systematic and structured approach is required to identify and quantify all cost components of the development and application of the tested evaluation methods, and to develop a conceptual framework for the qualitative assessment of the effectiveness of the tested indicators and evaluation methods, applied in the public good case studies in WP6. While the bulk of the required information was directly collated through the project team, stakeholder interviews and workshops (as well as sessions at the annual SRG meetings) with representatives from monitoring organisations and evaluators were required to collate information on monitoring programmes and to validate cost and effectiveness assessments of the project team.

The report outlines the conceptual framework developed for the cost-effectiveness assessment, summarises results of the cost and effectiveness assessment of the evaluation approaches tested in the case studies paying particular attention to the implications and role of different stakeholder priorities for the assessment of the effectiveness of evaluations, and provides synopsis of the implications of different data and monitoring programme scenarios on the cost-effectiveness of the tested evaluation approaches.

2 Concept and Scope of the Cost-effectiveness Assessment

The cost-effectiveness assessment in the ENVIEVAL project covers the steps of the evaluation cycle, namely the evaluation design, data generation and monitoring, database development, the application of evaluation methods and the interpretation of the results. The five main phases of the evaluation cycle are linked with the steps of the logic model framework that was developed within the ENVIEVAL project. This helps to link the cost and impacts on the effectiveness of an evaluation approach to each activity related to the planning, application and interpretation of the

method. Detailed information on the logic model framework can be found in the methodological handbook, the Deliverables D3.3, D4.3 and D5.4 as well as in the summary report of the case study testing (D 6.3) which provides examples of the logic model application.

The following schematic diagram provides an overview of the steps needed for the cost-effectiveness assessment of each evaluation method tested in public good case studies.

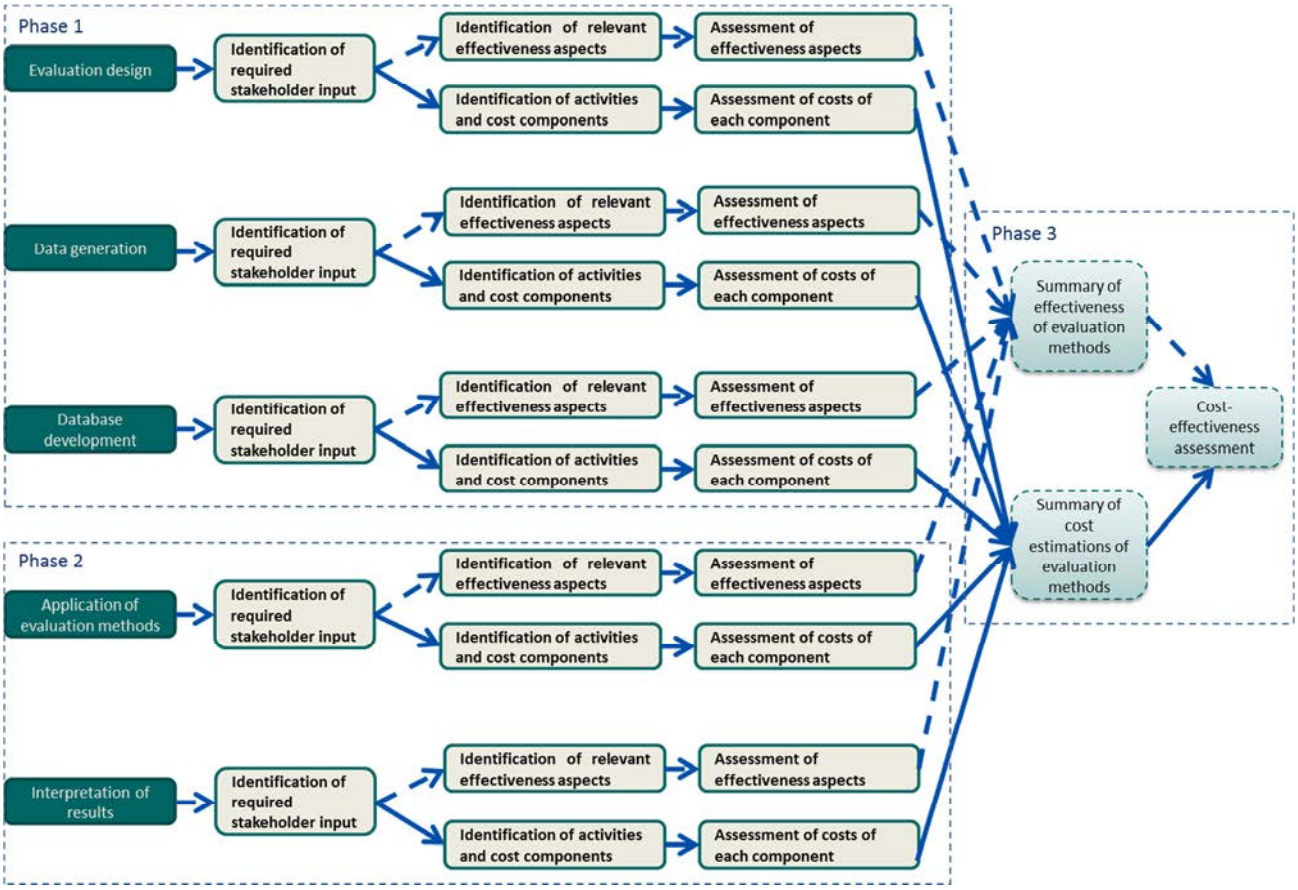

Figure 1: Schematic overview of the cost-effectiveness assessment

For each of the five main phases of the evaluation process, data on the costs of carrying out those steps and their contributions to the improvements of the effectiveness of the evaluation methods (e.g. in terms of the quality and robustness of the results) need to be collated. In Phase 1, the data collation and assessment was carried out for the evaluation design, data generation and the development of the databases in the public good case studies. In Phase 2, data on the cost and effectiveness aspects for the application of the evaluation methods as well as the interpretation of the results were collated during the case study testing in WP6. The collated data was then incorporated into the cost-effectiveness assessment of the evaluation methods in the third phase, carried out in WP7.

The cost-effectiveness assessment builds as much as possible on the experience and data from the case study testing (i.e. for example with respect to the database development and the application of the evaluation methods). Generally, the assumption is thus that, except for data on the cost and effectiveness aspects of monitoring requirements and programmes, data can be collated through the project team and does not require substantial input from external sources. However, at the beginning of the data collation exercise, each partner reviewed and identified to what extent the data for the cost and effectiveness assessment can be provided by the project team and to what extent input from stakeholders is necessary. Interviews with representatives of monitoring organisations were required to collate data on costs of monitoring programmes of environmental indicators and their potential impacts on the effectiveness of the evaluation method. Information on evaluator experiences with currently applied evaluation methods (e.g. use of data sources) was collected in the stakeholder consultation in 2013 and reviewed by the partners for this assessment.

In addition, stakeholders were consulted for validation at different points in time during the cost-effectiveness assessment. The first validation exercise took place at the SRG workshop in Budapest at the beginning of July 2014. Evaluators were consulted on the results of the cost assessment of the first three phases on the design of the evaluation procedures, data generation and database development as well as on the concept for assessing and comparing the effectiveness of the evaluation methods. The second validation exercise focused on the assessment of the effectiveness of the tested evaluation methods by using judgement criteria. The approach was validated with stakeholders in national stakeholder workshops during the case study testing of the evaluation methods. A third major validation exercise was conducted during the third stakeholder workshop in Vilnius in June 2015. The synopsis of the national stakeholder workshops was presented to the SRG members and they were asked to review and validate the approach of the performance assessment. For further validation, the performance assessment of each case study was validated at the final project meeting in September 2015. Each partner presented the performance assessment of their case studies to the partners and had to justify the performance level of each judgement criterion.

Following the identification of the required involvement from stakeholders into the data collation, the different activities (development of the evaluation design, to generate data, to develop the database and to apply the evaluation methods and interpret the results) and their underlying cost component were identified and then quantified. Generally, two perspectives need to be considered in the cost assessment. First, cost that arises for the evaluator in applying the evaluation method

(including set up of evaluation procedures, required data generation and database development) should be assessed. This assessment helps to inform evaluators in the selection of a cost-effective evaluation approach within the budget constraints of their evaluation contracts. Second, the overall cost of the evaluation method, including monitoring cost incurred by monitoring organisations and ministries, have to be assessed. This also takes into account that, when monitoring data are available free of charge for the evaluator, the cost need to be attributed to the evaluation method to get an holistic view on the overall cost-effectiveness of the evaluation method from a societal / taxpayers point of view.

The cost assessment seeks to determine the absolute and relative magnitudes of the cost of the evaluation method and the main determinants of high and low relative cost. The analysis of the main determinants is based on a qualitative analysis to identify the characteristics and structural factors that are common to evaluation methods and approaches with high or low relative cost. Some general rules and key issues were drawn for the identification and quantification of the main cost components such as labour, consumables, travel cost, indirect cost and transaction cost for the different tasks.

Similarly, relevant aspects for the effectiveness assessment need to be identified and assessed. The effectiveness of the evaluation approaches is defined as the performance of evaluation approaches to address the main evaluation challenges identified at the beginning of the project and thus to increase the effectiveness of evaluations.

Standardised criteria will be employed to harmonise the categorisation of the performance of the various methods into levels. Building on the Network of Networks on Impact Evaluation (NONIE) (2009), the CMEF requirements and the currently emerging guidelines for the ex-post evaluations refer to quality criteria such as: credibility, rigour, reliability, robustness, transparency, validity and practicability. While those criteria are well documented as theoretical quality aspects for a performance assessment of indicators and evaluation methods, the complexity of the seven qualitative criteria might constrain their practicability and acceptance in the stakeholder consultations and their usefulness for the methodological handbook. A less complex approach can be derived from a set of criteria developed by the EC (2001) to assess indicators to monitor the integration of environmental concerns into the CAP. This approach suggests the following quality criteria: policy relevance, responsiveness, analytical soundness, measurability, ease of interpretation and cost-effectiveness. Since policy relevance can be assumed as a given and cost-effectiveness is the planned outcome of the assessment, this leaves the four quality criteria -

responsiveness, analytical soundness, measurability, and ease of interpretation for the performance assessment. This approach was also used successfully in previous projects such as the agri-environmental footprint project.

The conceptual contribution of the ENVIEVAL project is to translate this approach into the context of assessing evaluation methods (instead of policy measures) and to integrate it into the performance assessment the degree up to which the main evaluation challenges are dealt with. The elimination or reduction of the challenges is an important quality criteria and the assessment of the cost-effectiveness of the new evaluation methods needs to explicitly consider the impact of these methods on reducing the evaluation challenges.

The evaluation challenges are integrated as judgment criteria for the four different quality criteria of the performance of the evaluation methods. The quality criteria for the assessment are categorised into levels of an ordinal scale. This requires the specification of a set of judgment criteria defining the different impact levels at ordinal scale such as low, medium or high (see for example Faehrmann and Grajewski, 2013). Figure 2 summarises the quality criteria, the suggested judgment criteria and impact levels for the performance assessment of the methods and indicators.

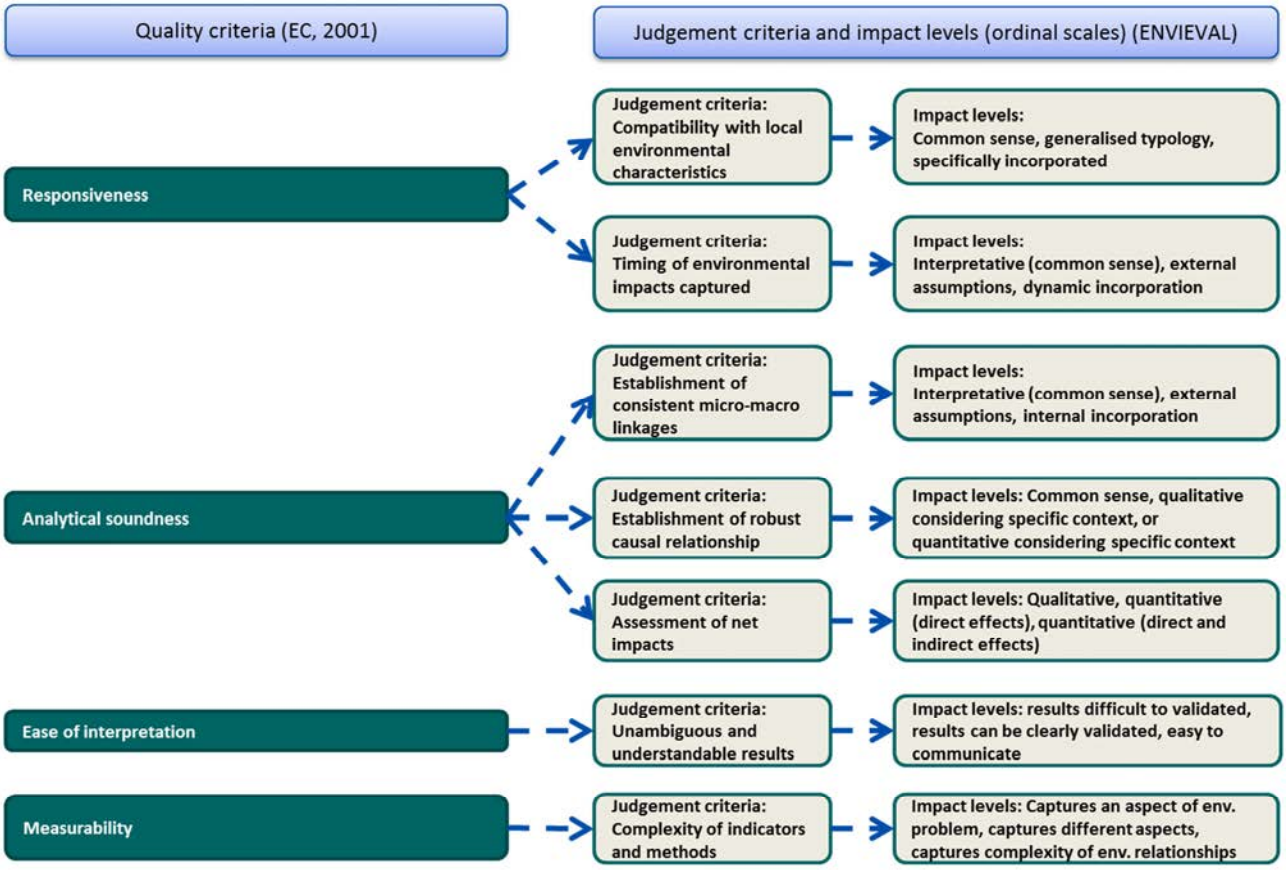

Figure 2 Overview quality criteria and impact levels

The differentiation of the impact levels in Figure 2 reflects ordinal scales of low, medium and high for each judgement criteria. The different impact levels are explained in Table 1.

Table 1: Definition of different performance levels

Judgement criteria	Impact level	Explanation of impact level
Compatibility with local environmental characteristics	Low	Applicability at the local level is assumed, common sense models are used to consider local environmental characteristics in the interpretation of the results.
	Medium	A generalised typology of environmental characteristics is used. Local characteristics are placed within this typology and addressed accordingly.
	High	Local environmental characteristics are specifically considered and incorporated into the evaluation approach.
Timing of environmental impacts captured	Low	Temporal dimensions of environmental impacts are not incorporated and only considered in the interpretation of the results
	Medium	Temporal dimensions of environmental impacts are incorporated in the methodological / evaluation approach through external assumptions
	High	Temporal dimensions of environmental impacts are directly incorporated in a dynamic modelling framework
Establishment of consistent micro-macro linkages	Low	Consistent micro-macro consistent only intuitive
	Medium	Incorporation of external assumptions to ensure (as much as possible) consistency between micro and macro level results
	High	The methodological approach explicitly covers and combines micro and macro-level analysis. Consistency and validation procedures are internalised.
Establishment of robust causal relationships	Low	Common sense models are characterised by relatively weak formulations of relationships that are not clearly evidence-based, but rather reflect general perceptions of how environmental outcomes are linked to interventions. Internal and external validity are not ensured.
	Medium	Qualitative models and methods are based on theoretically sound evidence but are not able to predict effects in quantitative forms.
	High	Quantitative approach based on well-documented, theoretically sound models and methods that facilitate robust prediction of how policy measures induce changes in agricultural practices which affect specific environmental issues. Internal and external validity are ensured.
Assessment of net-impacts	Low	Assessment of environmental net-impacts done in qualitative way highlighting only directions of impacts. Indirect effects are not considered.
	Medium	Methodological approach quantifies policy impacts but includes only direct effects
	High	Methodological approach quantifies net-impacts including direct effects and relevant indirect effects
Appropriateness of indicators and methods to capture complexity of environmental relationships	Low	Indicators and methods are linked to an aspect of the environmental problem / public good and are responsive to policy induced changes of agricultural system
	Medium	Indicators and methods reflect impact on different aspects of environmental quality / trade-offs
	High	The tested approach captures the complexity of environmental relationships and delivers measurements of change and impact of the relevant indicators.
Unambiguous and understandable results and policy recommendations	Low	Substantial efforts are required by the evaluator to translate the outcome of the application of the evaluation into understandable results and policy recommendations.
	Medium	Results are unambiguous and require little effort to be translated into understandable policy recommendations.
	High	Results are unambiguous and are easily translated into understandable policy recommendations.

The final task of the cost-effectiveness assessment is to conduct a cost-impact synopsis to compare and assess the cost-effectiveness of the tested indicators and evaluation methods and to inform the fact sheets for the methodological handbook in WP8. The qualitative cost-performance synopsis seeks to achieve a synthesis of costs and performance of methods and indicators. It helps to illustrate the structure of the total costs of the evaluation tools and the proportion of costs attributable to the different performance levels. By the integration of the cost-effectiveness aspects in the methodological logic model framework, the main decisions in the evaluation process and their influence on the cost and effectiveness of the tested evaluation approach are identified. Further, possible solutions for dealing with data gaps and the related impacts on the cost-effectiveness of evaluation approaches are tested using selected examples of the case studies. The selected case studies developed scenarios related to the improvement of the performance of the evaluation approach usually by improving data availability or using additional data sources. Taking into account the results of the analysis of the determinants of the cost, the synopsis will assess the proportionality or disproportionality of the cost-effectiveness balance of the new evaluation tools in the context of different circumstances, considering for example differences in data availability and skills.

3 Methodological Approach and Data Collection

3.1 Approach of the cost assessment

The main aim of the cost assessment is to analyse the importance of the different cost components and the determinants of the costs of the ENVIEVAL case studies, and to test a systematic approach for the comparison of costs of applying different methods in environmental evaluations. This assessment builds on cost templates that were completed by the project partners for their respective public good case studies of the ENVIEVAL project. In some cases, collaboration with evaluators and monitoring organisations was necessary to be able to complete the information for the monitoring cost.

The main steps of the task were to identify the required cost components in collaboration with WP6 (cases studies). The value or cost of the components and the overall costs of the developed evaluation tools was determined in absolute and classified in relative terms. An analysis of the main determinants of the costs was carried out using the experience from the testing of the developed evaluation tools in WP6.

The case studies cover seven public goods such as water quality, soil quality, climate, landscape, biodiversity HNV and wildlife, and animal welfare in case study areas in seven member states. The approach of the cost assessment follows the five main evaluation phases identified in the project (see Figure 1) and the main cost components as determined in the structure of the cost templates. Relevant cost components need to be identified and quantified for these relevant steps of the evaluation process.

3.1.1 Activities in the evaluation process and their cost

The cost of the evaluation approaches is collected according to the different phases of the evaluation cycle and the related activities that are attributed to the steps of the logic model framework. Therefore, the cost template includes all logic model steps that are attributed to the five phases of the evaluation process.

Design of the evaluation procedures

One of the first steps of the evaluation process is to establish a clear understanding of the evaluation task as well as to identify data requirements for the evaluation methods. The design of the evaluation approach has to be developed; this includes the identification of relevant indicators, including CMES and additional environmental indicators. Further, data requirements and available data need to be reviewed. The common unit of analysis has to be selected and conceptual decisions on counterfactual micro and / or macro level evaluations have to be taken. The following figure presents an overview of the tasks related to the evaluation design.

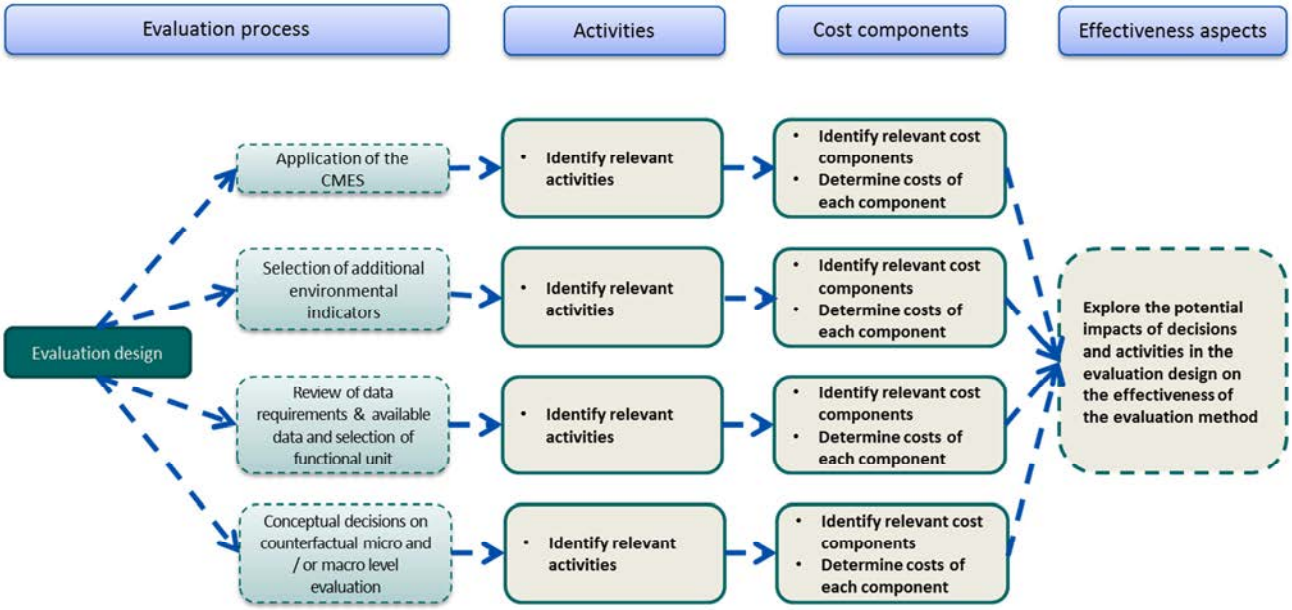

Figure 3 Overview of the task related to the evaluation design

- Application of the CMES

Evaluators are facing the tasks of assessing the environmental impacts of the programme and different relevant measures. Starting with the formal requirements and general intervention logic of the Common Monitoring and Evaluation System (CMES), evaluators need to select relevant measures of the RDP and evaluation questions for the environmental objective(s) they want to evaluate the measures / programme against. Then, output, result and impact indicators need to be selected and reviewed in the context of the available data. Tulloch et al. (2011) developed and evaluated approaches for a cost-effective and useful indicator selection. Selection criteria such as the indicator being easy-to-measure or the historic prevalence of data, are often more important than considering responsiveness to management. Also, these criteria have a high impact on cost-effectiveness. A better understanding of the different cost components could facilitate the use of more elaborated indicators. By including information about monitoring costs, leverage, certainty, benefits and probability of management success, the indicator selection process could increase the efficiency and effectiveness (Tulloch et al., 2011).

It is recommended to also consider at this stage to what extent the available data will later on enable the coverage of unintended effects on the environment and indirect effects such as deadweight and leverage effects at micro level and substitution and displacement effects at macro level. The inclusion of indirect effects into the evaluation design requires sufficient available data for non-participants.

The main costs associated with these tasks are related to working hours of the staff or capacity building through training activities and thus have an impact on labour cost. Other costs are usually not necessary for the conduction of these tasks.

- Selection of additional environmental indicators

While the CMES provides useful guidance on the general intervention logic, the number of environmental impact indicators is limited and, depending on the public good and environmental objective against which the measure and programme is evaluated, it becomes necessary to identify and select more suitable indicators to quantify environmental changes and to establish robust causal relationships between the policy-induced land management (or livestock management in the case of animal welfare) changes and measured environmental change. The suitability of the selected indicators needs to be reviewed in the context of their data requirements and the available environmental monitoring data.

As highlighted in Step 1.1, it is recommended to also review at this stage to what extent available environmental monitoring data cover non-participants and will later on enable the coverage of unintended effects on the environment as well as indirect effects such as deadweight effects at micro level and substitution effects at macro level.

- Review of data requirements and available data and selection common functional unit.

Depending on the public good and environmental objective, the selected indicators and available data as well as the level of analysis (micro or macro), a common functional unit applied to all used data needs to be defined for micro level and macro level evaluations. The functional unit (FU) can be defined as the 'smallest part of an organized system' (parcels, farm as agro-ecosystem, landscape unit, ecological area, sub-catchment area, etc.). The FU refers to the unit of study for assessing functional contributions of a system under a specified metric and delimits the analysis and the comparison of the organised system. Furthermore, the FUs are characterised by homogeneous activities and allow solving the scale interdependencies which is an important aspect to be defined for the logic model implementation. Examples of common functional units include farm (micro), catchment and regional units (macro).

- Conceptual decisions on counterfactual micro and / or macro level evaluations

Counterfactual based micro level evaluations are then designed and possible aggregation or upscaling of micro level data and results to macro level are reviewed. Alternatively, a separate counterfactual-based evaluation design is developed for macro level assessments. In either case, consistency checks between micro level and macro level results are required.

It can be concluded that costs in this phase of the evaluation cycle are mainly related to additional staff time spend for the development of the evaluation design. Decisions in this phase have strong impacts on the conceptual soundness of the evaluation design and on the overall outcome of the evaluation, and thus also on overall cost-effectiveness of the evaluation approach.

Data generation

As part of the data generation phase, existing primary and secondary data need to be reviewed and, if necessary, additional primary data collected, which can either be collected by the evaluators (e.g. through farm visits, surveys and questionnaires) or through monitoring programmes of environmental indicators. This serves to build a comprehensive data basis for the further evaluation steps.

There are two threads of costs that need to be considered in this assessment: First, the costs that arise for the evaluator and thus for the application of the evaluation method in itself should be

assessed. This analysis should provide information on the cost of the application of the tested evaluation methods for the evaluation organisations. Second, the overall cost of the evaluation method, including monitoring costs, has to be included. Also when the monitoring data can be accessed for free by the evaluator, the costs need to be attributed to the evaluation method to get an holistic view on the total cost of the evaluation method. If monitoring data do not exist, the cost of data collection has to be attributed to the evaluation method as well.

The collation of the data on cost and performance aspects requires input from the project teams and stakeholders (evaluators and monitoring organisations) in each partner country. The relevant activities and related cost components need to be identified and determined for each main task of this part of the evaluation process. In addition, potential impacts of the additional primary data on the performance of the evaluation process and methods need to be explored. Even if the use of primary data is not foreseen to test the methods in a public good case study, interviews with the organisation that is conducting the data collation can still explore the costs and potential benefits of existing and additional monitoring data for the evaluation process.

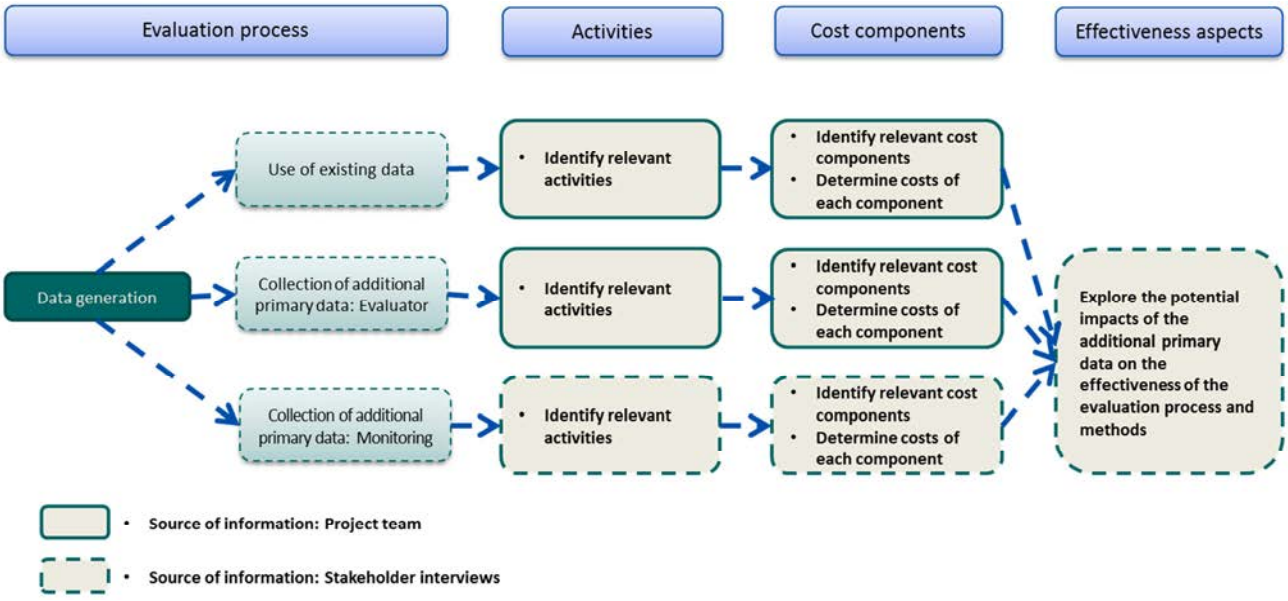

Figure 4 Overview of the required tasks for the data collection

Activities related to data generation can be attributed to step 2.1 of the logic model.

- Use of existing data (review and revision)

After identifying the data needs, it is required to check which existing primary and secondary data is available and if adjustments are necessary, e.g. by additional data collection. If existing data sets are exploited, important activities are to ensure access to it as well as the data preparation activities such as clearing and merging. Cost of these activities increase with the complexity of tasks and

data sets, and with the number of institutions involved in monitoring and data management. Furthermore, legal aspects, such as purpose of data and data confidentiality requirements, increase efforts to establish access to existing data. Qualified staff with matching expertise is important for the set up and further steps of the evaluation. Cost components include labour and personnel, equipment and consumables, and transaction cost. If existing monitoring serves the purpose of evaluation, the cost involved (or share of cost) should be considered in the cost-effectiveness assessment. This is important as in some cases data might be available, while in other regions no such monitoring exists, or access to data is denied. For these cases, we need information how much this additional monitoring would cost.

- Collection of additional primary data by the evaluator

If collection of additional primary data by the evaluator is necessary, the methodology and tools for the data collection have to be well defined and established. Tools for qualitative and quantitative analyses include interview guidelines and questionnaires, and guidelines for case study areas. The method of data collection (e.g. farm visits, surveys, questionnaires, sampling strategy, etc.) strongly influences its cost. Costs for conducting the data collection appear for the preparation of data collation and the implementation of pre-tests as well as the performance of sampling or surveying itself. Training on data collection methods or use of equipment might be necessary. The data collection includes fieldwork and laboratory work. This could be associated with a high labour demand and requires suitable equipment and facilities.

To assess and compare the cost-effectiveness of four different biodiversity indicators, Targetti et al. (2011) included the cost components staff time, distance and duration of travel, consumables and equipment in the analysis. They noticed that the largest share of costs is attributed to the field work and analysis of the samples, with labour being the main cost source. Desk and laboratory work are only a small part of the total cost. The four analysed indicators vary in costs mainly due to different duration of sampling. Good organisation of data collection as well as the use of cheap labour force (e.g. student workers) could reduce the monitoring costs (Targetti et al., 2011). The Hungarian biodiversity wildlife case study using the Farmland Bird Index depends on the commitment of volunteers for data collection. However, it is important to mention that these are very specific conditions of the studies as student workers and volunteers usually cannot be employed for the evaluation of RD programmes. Therefore, the use of cheap labour force or volunteers should not be recommended as it does not meet the reality of RDP evaluations. As

monitoring frequently requires highly qualified, specialised personnel, trade-offs between reducing labour cost and the quality of monitoring have to be considered.

The adequate planning and design of sampling has an effect on the cost and effectiveness of data collation. Carlson and Schmiegelow (2002), Lindenmayer et al. (2012) and the evaluators of environmental impacts of the Scottish RD programme (FERA, 2009) tested and compared different sampling strategies with regard to number of sample sites, frequency and detail of sampling. While Carlson and Schmiegelow (2002) conclude that it is more cost-efficient to monitor a larger number of sample sites more infrequently than a smaller amount of sampling sites with a high frequency (sampling frequency is costly and a larger number of sites increases the database), the other two studies suggest that a smaller number of sample sites with a more detailed data collection is more cost-effective. The impacts on the cost-effectiveness of different sampling strategies were tested and compared in scenarios of improved monitoring programmes and data availability of selected public good case studies.

The collection of additional primary data is linked to specific objectives and expectations how the additional data would improve the evaluation process and results. A qualitative judgement of the expected benefits and impacts of the additional data on the effectiveness of the evaluation process was considered by each partner in the cost-effectiveness assessment. The assessment was discussed with the national stakeholders and the stakeholder reference group at the stakeholder workshop in Budapest in July 2014.

Even if additional primary data are not collated in the public good case studies, it would still be useful for the cost-effectiveness assessment to collate information on what costs and expected benefits the specifically identified additional primary data would generate, if these data would be collected. This was tested in the developed scenarios of improved monitoring programmes and data availability in some of the case studies.

- Collection of additional data by monitoring organisations

Available monitoring of environmental data is often not sufficient to provide a comprehensive data base for evaluation of EAFRD measures. A typical shortcoming is that external monitoring data do not provide sufficient cases of beneficiaries and comparable non-beneficiaries. As such data are a pre-condition and starting point of M&E activities, cost of monitoring for the data used in the case studies will be assessed. If monitoring is performed with the purpose of evaluating specific measures, the full cost can be attributed to the M&E activity. However, in many cases monitoring

serves different purposes (e.g. advice, measurement of local environmental status), so that costs have to be shared between different uses.

From the point of view of the individual evaluator, it might be necessary to purchase data sets from monitoring organisations which could be very costly. It often depends on whether the evaluator is associated with a private-sector or public institute as the latter could facilitate access to data. Information on the cost of purchased data could be obtained from past transactions of the partner institutes, if data purchasing is not necessary for the case study testing. Price lists available from monitoring organisations also could, at least in some cases, provide an estimation of the real cost of the monitoring activities. The quality and quantity, as well as the type, of monitoring data available affect the effectiveness of evaluation methods. Hence it is important to include costs and potential impacts of monitoring programmes on the effectiveness of evaluation methods in our cost-effectiveness assessment.

A well-defined and well organised set up of the data generation and monitoring process would increase the expected benefit and effectiveness. Monitoring activities should be targeted to the evaluation questions while unnecessary data generation should be avoided. As it is often difficult to allocate costs to the quality of monitoring and the effectiveness of evaluation methods, at least a qualitative judgement of these attributes should be carried out.

As a source of information, selected stakeholders of monitoring organisations were interviewed to provide information on the costs of the monitoring programmes (for indicators relevant to the respective public good case studies). They provided their expert judgement on the potential benefits different monitoring efforts might have for the quality and the analytical soundness of the evaluation results, as well as regarding other purposes of the monitoring efforts.

Database development and maintenance

Costs related to the development of a database are associated with the construction as well as the maintenance of the database. In this assessment the focus will be on the databases generated for the case studies in the ENVIEVAL project. Therefore the partners were able to integrate the assessment of the cost from the beginning of the development of the case study databases. This could provide exact data for the cost assessment.

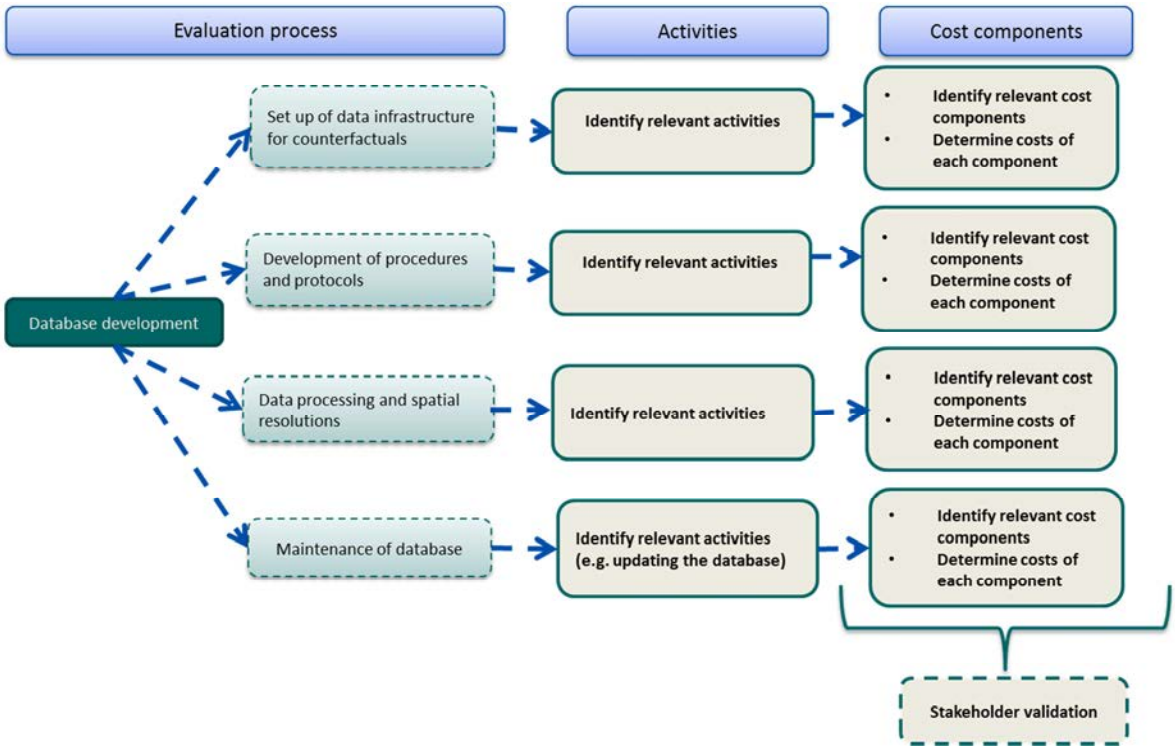

Figure 5 Overview of the required tasks related to the database development

- Database development, Step 2.2, 2.3, 3.1, 3.2, 4.1 and 4.2

The development of the database by the evaluator includes activities such as the set-up of a data infrastructure for counterfactual analysis including data formats and data rights (Steps 2.2 and 2.3). Further, the development of procedures for dealing with different demarcation of geographical units between different data sets, and protocols for aggregating and anonymisation of individual farm and firm data need to be considered (Steps 3.1 and 4.1). Ensuring storage and securing data is also an important part of the database development.

- Maintenance of the database, Step 3.2 and 4.2:

To enable continuous and up-to-date access to, as well as improvements of, the database, the maintenance and further revisions have to be provided. In this process, important activities could be the programming and updating of the database which requires programming and informatics skills. Thus, the main source of costs is associated with experienced and skilled staff and the required equipment and facilities such as soft- and hardware, as well as transactions for getting access to new or updated data sets.

The costs of the database development increase with its complexity and the multitude of data sources. This aspect should be linked with the expected benefits and impacts on the effectiveness of the database development. The set up and maintenance of a reliable database is important for

the effectiveness of evaluation and monitoring programmes. If collected data remains unused or is even lost due to inadequate documentation and quality control, the cost-effectiveness of monitoring and evaluation will decrease. Testing of different degrees of database complexity/different amounts of data sets allows the comparison of cost and effectiveness and detect key points for optimisation and improvements.

The experiences of the project partners with the implementation of the database for the public good case studies build the basis for the assessment of the cost components of this evaluation step. Therefore, the costs of the different cost components were determined during the development of the case study databases. Although the main information will be provided by the project partners, stakeholders were consulted for validation of the costs allocated to the different activities and components. Only the development of databases linked to the evaluation methods were considered. General databases, e.g. of monitoring data established by the monitoring organizations were attributed to the cost of monitoring (see 3.1 Data generation and monitoring) and not included in this evaluation step.

Application of the method, Step 3.3 and 4.3

This evaluation step includes the implementation of the evaluation methods (e.g. development of the modelling framework) and the conduction of the analysis of RD impacts. Activities such as the review of indicators and the development of the conceptual model as well as the execution of the analysis are relevant. Counterfactual analysis at micro and / or macro level has to be considered.

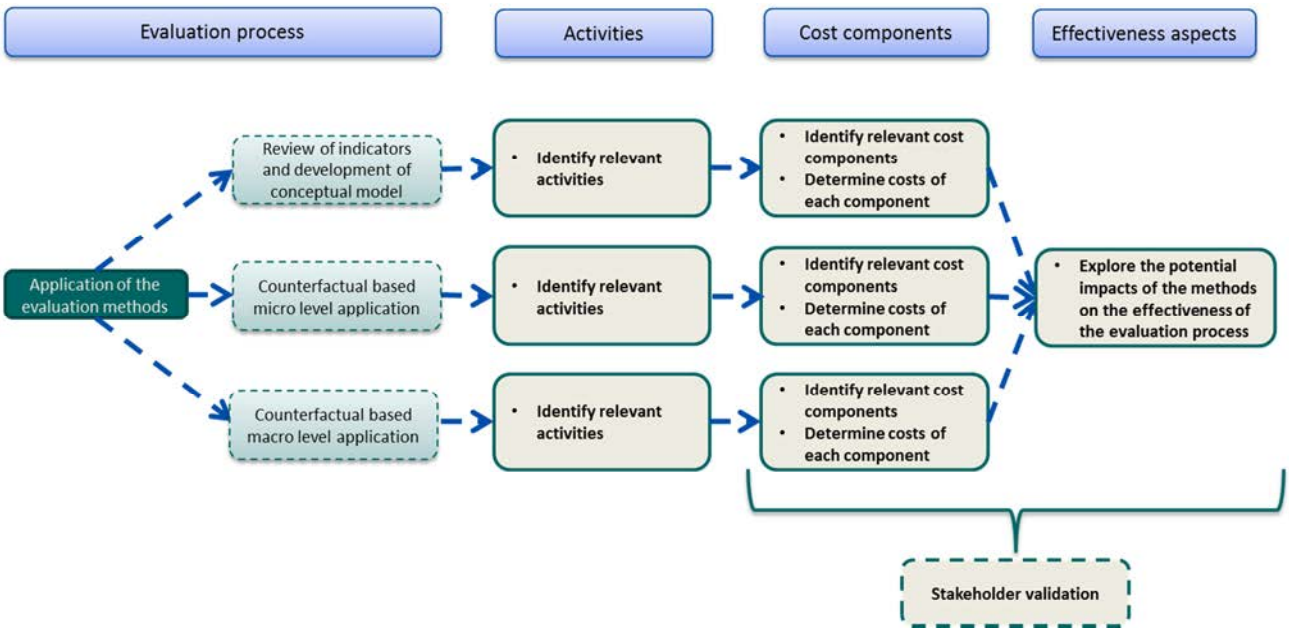

Figure 6 Overview of the required tasks for the application of the evaluation method

Depending on the selected evaluation methods at counterfactual, micro and macro level this can include the setting up of the modelling framework including the integration of indicators and data processing and synthesising of available data and information. Activities and cost components need also to be considered in relation to the actual application of the evaluation methods at counterfactual, micro and macro levels, for example the simulation of different scenarios with an economic modelling framework or the application of bio-physical and / or statistical models to assess net impacts. Upscaling activities could be necessary to analyse impacts at the programme level. The extent of the cost of this evaluation step largely depends on the complexity of the selected methods to be tested in the case studies. Finally, cost components (e.g. required staff time and qualification) for the interpretation of the evaluation results and the assessment of RD impacts need to be considered.

To complete these tasks, qualified staff and relevant equipment (computer programmes) are demanded. The main cost component in this step is presumably labour and personnel. The costs depend on the complexity and completeness of the selected methodological framework and available data. Differences of methods that have an influence on the cost should be emphasised to facilitate the comparison of different evaluation methods. Effective and resource-efficient application of the evaluation methods is necessary to optimise cost-effectiveness. The expected benefits of the evaluation methods and impacts on effectiveness are also a component of this assessment. The experiences of the project partners with the application of the tested evaluation methods in the public good case studies build the basis for the analysis of the cost components of this evaluation step. Although the main information was provided by the project partners, stakeholders were consulted for validation.

Interpretation of results of RD impacts and consistency checks, Step 3.4 and 4.4

Cost for the interpretation of results and RD impacts of the analysis and the development of sound policy recommendations have to be included for this evaluation step. Further, the application of consistency checks might be necessary as multiple data sources are required for micro-macro level evaluation, deriving from different databases and providing for different data with different metrics and terminology. The main purpose of the consistency checks is to ensure that the results of the policy impact analysis at micro level and macro level coincide.

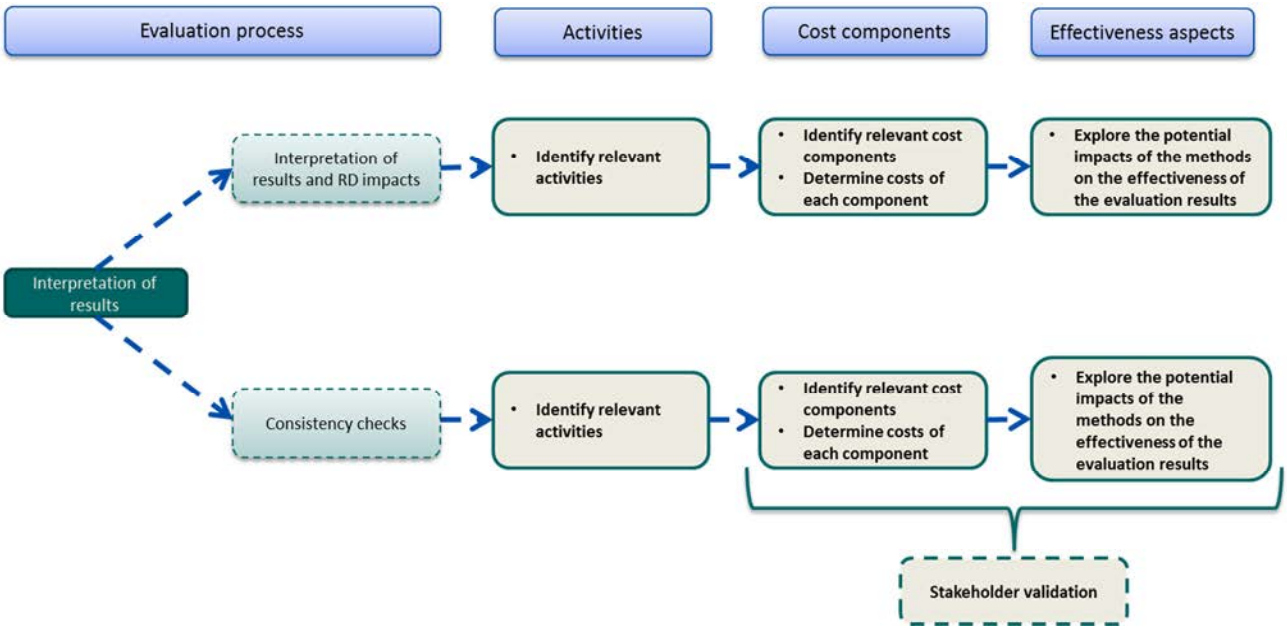

Figure 7 Overview of the required tasks for the interpretation of the evaluation results

The interpretation of the results and RD impacts creates labour costs which vary depending on the complexity of the results. The translation of the results into sound policy recommendations is also an important activity that needs to be considered for the cost assessment. The conducting of consistency checks is also mainly related to an increased work load for the evaluator and thus influences labour cost. The cost of these activities might be rather low, but they are essential for the successful application of the evaluation approach. The extraction of the key messages and to communicate them to the target group has a big influence on the perception and use of the RD impacts.

3.1.2 Main cost components

Relevant cost components need to be identified and quantified for these relevant steps of the evaluation process. Standard cost components are: a) Labour and personnel (considering required skills), b) Contracting cost, c) Equipment and other consumables, d) Travel cost, e) Indirect cost and f) Transaction cost. Some general rules apply for the quantification of the cost components:

- If data on labour cost are not available, estimations should be calculated using hours and number of staff attributed to this task, and approximate salaries considering the profession of the staff. Time sheets could be used to approach the amount and allocation of staff hours to different steps of the evaluation process. If time sheets are not available, similar ways of estimating the labour cost need to be addressed.

- Cost of training or consultation for evaluation methods (internal and external) could be approached by using the fee of the expert trainer as well as the hours of staff time during the training.
- Contracting cost include all cost related to the contracted activity and are not included in the calculation of indirect cost.
- Travel cost to case study areas could be determined by using invoices and total expenses of the field trips.
- Inclusion of indirect costs (overheads) such as rent, electricity, maintenance, taxes, insurance in the assessment is necessary. Indirect cost should be calculated according to the accounting rules of the partner organisation and, in addition, as a 25% flat rate to allow for comparison of the cost structures without bias due to different overhead rules.
- Transaction costs (e.g. administrative costs) of data acquisition need to be included in the assessment.

The collection of detailed information on each single cost aspect (unit price) enhances the possibility to select the most cost-effective allocation of costs (e.g. number of samples, study sites, and staff hours) and helps to illustrate the structure of the total costs of the evaluation tools and the proportion of costs attributable to the different performance levels. Potential impacts on the performance of the evaluation methods need to be identified and assessed qualitatively in relation to the quality criteria (see Table 1).

3.2 Approach of the performance (effectiveness) assessment

3.2.1 Framework for the performance assessment

A standard feature of multi-criteria and cost-effectiveness assessments is a performance matrix, or consequence table, in which each row describes a methodological approach tested in the case studies and each column describes the performance (here used as a synonym for effectiveness) of the approach against each criterion. Table 2a provides an example for the structure of a performance matrix for the ENVIEVAL effectiveness assessment.

Each partner team produced performance matrices for their case studies with different rows for the variations in the methodological approach tested. Variations can, for example, be the selection of different indicators or different methods to incorporate counterfactual analysis into the evaluation approach. For each of the variations (i.e. each row) the performance was assessed by allocating an impact level (low, medium or high) to each judgement criteria. These performance matrices of the

own case studies of each partner were discussed and validated at the national stakeholder workshops.

Table 2a Performance matrix with impact levels (example highlighting the structure)

Methodological approach	Performance criteria						
	Responsiveness		Analytical soundness			Ease of interpretation	Measurability
	Compatibility with local env. characteristics	Timing of env. impacts captured	Establishment of consistent micro-macro linkages	Establishment of robust causal relationships	Assessment of net-impacts	Unambiguous and understandable results	Appropriateness of indicator/method
Methodological approach 1	Low	High	High	High	Low	Low	Medium
Methodological approach 2	Medium	Medium	Medium	Medium	High	Medium	High
Methodological approach 3	High	Low	Low	Low	Medium	High	Low
Methodological approach 4	Medium	Low	High	High	High	Low	Medium

Qualitative sign analysis

Each partner has then carried out a qualitative sign analysis of the performance matrices. A qualitative sign analysis is a simple synthesis tool which is based on the assumption that qualitative data cannot be added up, but that the frequency of occurrence of an impact level over a range of criteria can be numerically treated (Nijkamp and Blaas, 1994). The number of high, medium and low scores of each methodological (evaluation) approach of the performance assessment in the performance matrix is synthesized in a frequency table. Table 2b provides an example for a frequency table.

Table 2b Frequency table of a qualitative sign analysis (numbers for illustrative purposes only)

Methodological approach	High	Medium	Low
Methodological approach 1	3	1	3
Methodological approach 2	2	5	0
Methodological approach 3	2	1	4
Methodological approach 4	3	2	2

Allocation of different weights

The performance matrix was expanded through the allocation of different weight (priority) categories to the judgement criteria by the stakeholders. The different weights – very high priority, high priority and medium priority - were applied in the stakeholder workshops. The reason for defining the three different weights as very high, high and medium - and not high, medium and low – is to use a more neutral wording for the lowest weight. The frequency of different scores can

then separately be summarised for each of three weight categories (i.e. for the criteria given a very high priority by the stakeholders, for the criteria with high priority and criteria with medium priority).

The starting point, however, was an equal weight for all judgement criteria. Care must be taken in the assessment at the level of the quality criteria as the different number of equally weighted judgement criteria under the four quality criteria will lead to different weights of the quality criteria, with higher weights or importance of those quality criteria with several judgement criteria. The allocation of different weights (i.e. priorities) to the judgement criteria was validated in the participatory stakeholder workshops asking stakeholders to weight the various criteria according to their importance for the effectiveness assessment of the different evaluation approaches. This assessment was done during the first part of the stakeholder workshop in each partner country. The participatory assessment provides a validation of the suitability and importance of the different judgement criteria from different stakeholder perspectives. The different weight allocations provide the basis to analyse and show how the outcome of the cost-effectiveness assessment of the evaluation approaches (and the resulting selection of the most cost-effective and suitable evaluation approach) can change depending on different stakeholder priorities.

The performance matrix and frequency table could already be the final product of – a rather simple – performance assessment, if more in-depth comparative analysis of trade-offs is not possible. The methodological handbook provides guidance on how to develop and use such a performance matrix and frequency table and the evaluators and end-users would then be left with the task of assessing the extent to which their objectives and circumstances are met by the results of the qualitative sign analysis derived from the evaluation approaches tested in the case studies. This would require a simple dominance analysis.

The qualitative sign analysis is an easily applicable multi-criteria assessment method for scoping the cost-effectiveness of alternative evaluation approaches. However, such rather intuitive processing of the data can be speedy and effective, but it may also lead to the use of unjustified assumptions, causing incorrect ranking of options. In analytically more sophisticated effectiveness assessments, the information in the basic matrix is usually converted into consistent and comparative scores and trade-off matrices are constructed and analysed. This implies that trade-offs between different criteria are acceptable, so that good performance of an evaluation approach on one criterion can in principle compensate for weaker performance on another. Where compensation is acceptable, most multi-criteria assessment (MCA) methods apply implicit or

explicit aggregation of each option's performance across all the criteria to form an overall assessment, on the basis of which the set of options can be compared. Since we only have qualitative data available for the effectiveness assessment, a qualitative outranking analysis would be a suitable approach for the ENVIEVAL effectiveness assessment.

Allocation of scores and trade off analysis

This part was mainly done by the TI team, building on the information from the different performance matrices and weight assessments at the stakeholder workshops. Detailed descriptions of a qualitative outranking analysis can be found as early as Van Delft and Nijkamp (1977). The approach is based on the assumption that performance on each criterion is categorised into one of four impact level (or categories) (●●●●, ●●●, ●● and ●) in descending order of quality. For our purpose, we suggest that the differentiation of three impact levels (low, medium and high) is sufficient and more practical for the participatory assessment with stakeholders. First, pairwise comparisons are made for each criterion between all pairs from among the methodological (evaluation) approaches being considered. Depending on the difference in assessed performance, each comparison might be: at one extreme a major positive difference (●●● (high) against ● (low)) coded as +2; at the other extreme, a major negative difference (● (low) against ●●● (high)) coded as – 2; or any of the intermediate assessments +1, 0, or –1. For each criterion, all the pairwise comparisons will be summarised by a skew-symmetric matrix, with zeros down the leading diagonal. The structure of the matrix is shown in Table 4 and 5 for the criteria 'Compatibility with local env. characteristics' and 'Assessment of net-impacts'.

Table 3 Skew-symmetric matrix for the assessment of the trade-offs between evaluation approaches (example highlighting the structure) - Compatibility with local env. characteristics

	Methodological approach 1	Methodological approach 2	Methodological approach 3	Methodological approach 4
Methodological approach 1	0	-1	-2	-1
Methodological approach 2	+1	0	-1	0
Methodological approach 3	+2	+1	0	+1
Methodological approach 4	+1	0	-1	0

Table 4 Skew-symmetric matrix for the assessment of the trade-offs between evaluation approaches (example highlighting the structure) - Assessment of net-impacts

	Methodological approach 1	Methodological approach 2	Methodological approach 3	Methodological approach 4
Methodological approach 1	0	-2	-1	-2
Methodological approach 2	+2	0	+1	0
Methodological approach 3	+1	-1	0	-1
Methodological approach 4	+2	0	+1	0

A concordance index c(i,j) will then be calculated, which represents the frequency with which option i is better than option j. Normally, the application of a concordance analysis and index requires the definition of weights for each criterion. In the first instance, we will treat the seven judgement criteria with an equal weight (priority) in the assessment of the case study results. The idea is then to apply the three different weights (priority) categories allocated by the stakeholders to each judgement criteria to analyse and show the differences in the outcome of the cost-effectiveness assessment of the evaluation approaches and the resulting consequences for the selection of the most cost-effective and suitable approach.

With an equal weight of 1 as the starting point of the assessment, only one concordance index will be calculated for each pairwise comparison. Once three different weight categories have been suggested by the stakeholders and allocated to each judgement criteria (taking into account the resulting weights for the four quality criteria), different concordance indices will be calculated for all criteria with the same weight. These outputs will be summarised in three concordance matrices. It is then possible to compute a net total dominance index by computing for the difference between the extent to which option 1 dominates all other options and the extent to which other options dominate option 1 (i.e. the sum of row 1 minus the sum of column 1) (e.g. Bouyssou et al., 2006, Van Delft and Nijkamp, 1977, DCLG, 2009). The structure of a concordance matrix is shown in Table 6.

Table 5: Example of the structure of a concordance matrix for the assessment of the evaluation approaches (numbers only for illustrative purpose)

	Methodological approach 1	Methodological approach 2	Methodological approach 3	Methodological approach 4	Sum of rows	Net total dominance index (net concordance index)
Methodological approach 1	0	0.43	0.57	0.71	1.71	**-0.15**
Methodological approach 2	0.57	0	0.71	0.71	1.99	**0.70**
Methodological approach 3	0.43	0.29	0	0.43	1.15	**-0.84**
Methodological approach 4	0.86	0.57	0.71	0	2.14	**0.29**
Sum of columns	1.86	1.29	1.99	1.85	*6.99*	

Vice versa, an unweighted discordance matrix can be calculated for each pair of methodological (evaluation) approaches in a similar way. The discordance index, d(i,j) is calculated as the frequency with which the outcomes of option i are much worse (-2) and slightly worse (-1) than option j. This information feeds directly into a discordance matrix D, from which in turn a net discordance dominance index can be computed, similarly to the concordance dominance index described above.

Final selection (in the cost-effectiveness synopsis) is not based on any fully defined procedure, but revolves around an inspection of the net concordance and discordance indices (at each of the three weight categories – priorities) seeking an option that exhibits high concordance and low discordance (especially with respect to the high priority criteria) considering the relative costs of the different approaches as well as specific circumstances, preferences and abilities of the end-user (stakeholder).

However, the feasibility of carrying out a trade-off and qualitative outranking analysis as described below depends on the available information and comparability of the evaluation approaches from the different case studies. The comparability of the tested approaches across the public goods is limited due to a different emphasis on particular evaluation challenges, micro and macro levels and certain evaluation approaches can only be apply in the context a specific public good. This implies that only a small number of approaches across the same or similar public goods can be compared. A small number of comparable approaches do not merit the application of a more in-depth outranking and trade-off analysis, but a qualitative sign analysis is used to highlight how different stakeholder priorities affect the interpretation of the results and ultimately selection of the approach for environmental impact evaluations of RDPs.

3.2.2 Participatory assessment and stakeholder validation

One of the main tasks and input into the development of a methodological framework for the assessment of the effectiveness of evaluations is the joint assessment with the stakeholders of the suitability and importance of the different quality and judgement criteria. At first, stakeholder workshops were conducted at national level of the ENVIEVAL partner countries. The national workshops took place between March and June 2015. Subsequently, the results of the national workshops were presented and discussed at the third international stakeholder workshop in Vilnius on June 10th and 11th with the stakeholder advisory group. Many SRG members also participated in the national workshops.

The two main objectives of the stakeholder workshops were:

- To validate the suitability and weights of the quality and judgement criteria of the performance assessment in different stakeholder contexts
- To review the assessment of the performance of evaluation approaches tested in the own case studies in the context of different stakeholder priorities for the various evaluation criteria included in the assessment framework
- To highlight and to create an understanding of the consequences of different stakeholder priorities for the design and the selection of evaluation approaches.

The participatory assessment of the suitability and weights of the quality and judgement criteria ensures the practical applicability of the selection of the different criteria for the performance assessment of RDP evaluation approaches and provides an overview of different stakeholder priorities concerning the different criteria included in the assessment framework. The use of different weights (priorities) delivers a more robust picture of the performance of the different evaluation approaches in the context of different stakeholder priorities with respect to the importance of different quality and judgement criteria.

The focus of the national workshops in the partner countries was on the context of the own case studies of each partner. The idea was to get a better understanding of stakeholder priorities for different evaluation criteria and how these different stakeholder priorities would affect the design and the selection of evaluation approaches. To highlight resulting differences in the stakeholder preferences for different evaluation approaches, one evaluation approach tested in another case study was briefly presented at the end of the workshop to show with a concrete example that the different stakeholder priorities can affect the selection of an approach. This was however just for an exemplary purpose. A detailed comparative assessment of the criteria and the performance of

the tested methods across case studies and partner countries were conducted at the international stakeholder workshop in Vilnius in June 2015.

The targeted stakeholders for the national workshops included evaluators, representatives from the managing authorities and monitoring organisations, i.e. 3 types of stakeholders with potentially different weights and priorities for different performance criteria. Ideally, about 5 participants per type of stakeholders attended the workshop (a total of 15 participants). We recognized, however, that it is difficult to find 5 participants per stakeholder type in smaller countries or countries with only one national RDP. In those cases 2 participants of each group still allowed to validate the suitability of the different criteria and to compare different weights and priorities between stakeholders. The background, expertise and interest of the participating stakeholders did cover both public good case studies. It was desirable for the comparative assessment in the international workshop in Vilnius that the SRG members also participated at the national workshops.

In preparation of the workshop each partner had to carry out a preliminary assessment of the performance of the evaluation approaches tested in their own case studies (that is to fill the performance matrix as outlined in Table 2 above). A short outline or background document briefly explaining the objectives of the ENVIEVAL project, the purpose of the workshop within the ENVIEVAL project and the specific objectives of the workshop as well as the key questions to be discussed were provided to the stakeholders with the invitation.

The workshops were divided into two main parts. The first part can be done independently from the results of the case studies, as the main purpose of this part is twofold:

1) to validate the methodological framework and evaluation criteria for assessing the performance of RDP evaluation approaches
2) to identify and define different stakeholder priorities (weights) for the evaluation criteria.

The second part of the workshop focused on the participatory application of the framework (and criteria) with the aims:

1) to review the performance assessment of the tested evaluation approaches with the stakeholders
2) to highlight the consequences of different priorities for evaluation criteria for the design and selection of the evaluation approach.

Methodological approach for the first part of the workshop

Following the introduction of the workshop, the framework and evaluation criteria (quality and judgement criteria) and the differentiation of the three impact levels for each criteria was explained. Table 2 provided the basis for the explanation. In addition to explaining the generic differentiation between the three different impact levels (as already included in Table 2), the criteria and their impact levels also were explained in a more concrete context of a specific public good. For example, what standard or characteristics would an evaluation approach to assess water quality impacts need to fulfil to gain a high impact level and not a medium or low impact level for the criteria 'robust causal relationship'? What kind of water quality aspects would the evaluation approach need to consider? A hand-out of the table was provided to stakeholders at the national workshops.

Further, the origin of the performance assessment framework (derived from a set of quality criteria developed by the EC (2001) to assess indicators to monitor the integration of environmental concerns into the CAP) was introduced to the workshop participants and the integration of the identified main evaluation challenges as judgment criteria for the four different quality criteria of the performance of the evaluation methods was explained to them. The purpose of the allocation of the weights and how the weights will be used in the performance assessment also was explained (i.e. differentiation between more important and less important criteria in the qualitative sign analysis and qualitative outranking analysis.

After the introduction and explanation of the framework the participants were divided into three break-out groups according to the different type of stakeholders. The advantage of this approach is that the evaluation criteria could be discussed separately with evaluators, monitoring organisations and representatives from the ministries. However, each partner reviewed the number and background of participants. If the number of stakeholders from monitoring organisations or ministries was too low (i.e. less than three) these stakeholders were merged into one group.

The first task in the break-out group was the validation of the framework for the performance assessment. Do stakeholders in principle agree with the integration of the main evaluation challenges as judgement criteria and are the established linkages between the evaluation challenges / judgement criteria and the quality criteria plausible for them? Do stakeholders fundamentally disagree with the inclusion of a specific criterion or do they feel an important aspect is missing and another criterion would need to be added?

- Is the presented concept of the performance assessment plausible and understandable?

- Does the framework cover all relevant aspects for the assessment of the performance of evaluation approaches?
- Does the framework cover the most important challenges in environmental RDP evaluations? If not, can you suggest of another quality and/or judgement criteria for the performance assessment of evaluation approaches?

In the next step, stakeholders were asked for an initial allocation of their priorities of the different judgement criteria.

Sheet 1: Exemplary structure:

- Please indicate your affiliation:
  - Evaluator ☐ Monitoring organisation ☐ Ministry / managing authority ☐
- Are you specialised in one or several particular environmental theme(s)? If yes, please specify:
- For how many years are you involved with RDPs and their evaluation?
  - 0 – 5 years ☐ 5 – 10 years ☐ longer than 10 years ☐

Judgement criteria	Allocation of dots
Compatibility with local environmental and farm structural characteristics	
Timing of environmental impacts captured	
Establishment of robust causal relationships	
Assessment of net-impacts	
Establishment of consistent micro-macro linkages	
Appropriateness of indicator(s) to capture complexity of environmental relationships	
Unambiguous and understandable results and policy recommendations	

Each stakeholder received a prepared sheet with a few questions on the background of the stakeholder and a table with the judgement criteria. Stakeholders were then asked to answer the questions and allocate 15 dots over the seven judgement criteria. The initial allocation of priorities (dots) across the judgement criteria was done without restrictions. Particular noticeable or unusual

allocations were picked up in the following discussion. In the context of the weighting system, the allocation of two dots for a certain judgement criteria represents a medium weight and high priority, while three or more dots represent a very high priority (highest weight) and one or less dots represent a medium priority (lowest weight). If stakeholders have identified another judgement criteria which was added to the framework, then the number of dots needed to be increased by two (17 dots for 8 judgement criteria).

The discussion in the break-out group started with exploring the reasons for different stakeholder priorities. Each stakeholder was asked to explain his/ her allocation of dots and to outline the reasons for allocating higher or lower priorities to certain criteria. The identified reasons provided an important input into the comparison of stakeholder priorities across the stakeholder types and partner countries. The discussion then focused on exploring with the stakeholders the potential consequences of their initial allocation of priorities to the judgement criteria for the desired performance of evaluation approaches. In other words, for which criteria would that suggest a higher priority of achieving a high impact (performance) level then for other criteria and what would that imply or suggest for the emphasis of the design of the evaluation approach? The discussion should also highlight how the allocated priorities of the judgement criteria affect the relative importance of the overarching quality criteria. The discussion should lead to a review of the allocation of the priorities (weights) with the aim of agreeing on a joint group allocation of weights, i.e. one, two or three dots for each criterion (which is equal to a weight or priority of medium, high and very high). If the participants were not able to agree on a consensus and joint group allocation of weights, averages were used to derive a group allocation. These weights were used in the second part of the workshop to review the assessment of the performance of evaluation approaches tested in the case studies in the context of different stakeholder priorities for the various evaluation criteria.

At the end of the first part a short feedback session was held where each break-out group reported back on the allocation of priorities (weights) to the different judgement criteria followed by a short reflection and discussion of the differences between the three groups.

To close the first part of the workshop each stakeholder was asked to fill in another sheet and to carry out a relatively simple pair-wise comparison of each criteria. The main advantage of this approach is that only two criteria are compared by the stakeholder at a time in terms of their relative importance. Pair-wise comparisons derive from the Analytic Hierarchy Process (AHP) developed by Saaty (1980) and originally apply a scale of 1 – 9 to differentiate between the

importance of two criteria. Since we use the resulting weights for qualitative methods such as the qualitative sign analysis based on simple ordinal scales we only applied the pair-wise comparison with a more important or less important differentiation. For each criteria pair, the stakeholder selected the judgement criteria, which is more important to her / him (see example table below). A similar approach has also been used in the BioBio project (Kelemen et al., 2011).

The pair-wise comparison provided more detailed information on the stakeholder priorities between the different criteria. The results of the pair-wise comparison were used – after the workshop - to validate the allocation of weights and to examine in consistencies in the allocation of priorities. These results were reported back to – and discussed with - the stakeholders at the international workshop in Vilnius.

Sheet 2 – Pairwise comparison of the judgement criteria

Which criterion is more important to assess the performance of an evaluation approach?

Please compare each pair of evaluation criteria in the table below and choose for each pair the criteria for which you think it is more important to achieve a high performance level in evaluations. Please add a 'X' in the column of the criteria with the bigger importance.

Criteria A	Criteria B	A	B
Timing of environmental impacts captured	Establishment of robust causal relationships		
Timing of environmental impacts captured	Assessment of net-impacts		
Timing of environmental impacts captured	Establishment of consistent micro-macro linkages		
Timing of environmental impacts captured	Appropriateness of indicator(s) to capture complexity of environmental relationships		
Timing of environmental impacts captured	Unambiguous and understandable results and policy recommendations		
Timing of environmental impacts captured	Compatibility with local environmental and farm structural characteristics		
Establishment of robust causal relationships	Assessment of net-impacts		
Establishment of robust causal relationships	Establishment of consistent micro-macro linkages		
Establishment of robust causal relationships	Appropriateness of indicator(s) to capture complexity of environmental relationships		
Establishment of robust causal relationships	Unambiguous and understandable results and policy recommendations		
Establishment of robust causal relationships	Compatibility with local environmental and farm structural characteristics		

Assessment of net-impacts	Establishment of consistent micro-macro linkages		
Assessment of net-impacts	Appropriateness of indicator(s) to capture complexity of environmental relationships		
Assessment of net-impacts	Unambiguous and understandable results and policy recommendations		
Assessment of net-impacts	Compatibility with local environmental and farm structural characteristics		
Establishment of consistent micro-macro linkages	Appropriateness of indicator(s) to capture complexity of environmental relationships		
Establishment of consistent micro-macro linkages	Unambiguous and understandable results and policy recommendations		
Establishment of consistent micro-macro linkages	Compatibility with local environmental and farm structural characteristics		
Appropriateness of indicator(s) to capture complexity of environmental relationships	Unambiguous and understandable results and policy recommendations		
Appropriateness of indicator(s) to capture complexity of environmental relationships	Compatibility with local environmental and farm structural characteristics		
Unambiguous and understandable results and policy recommendations	Compatibility with local environmental and farm structural characteristics		

Methodological approach for the second part of the workshop

Following a short introduction of the objectives of the second part of the workshop, the evaluation approaches tested in case studies and their prepared performance matrix were presented and explained. It was important to explain the advantages and disadvantages of the approach and how the approach tries to address the different main evaluation challenges. This was important background information on how the performance scores for the different criteria have been derived.

Table 6 Illustrative example for a filled performance matrix

Evaluation approach		Quality criteria						
		Responsiveness		Analytical soundness			Measureability	Ease of interpretation
		Judgement (performance) criteria						
Indicator	Method	Compatibility with local environmental … characteristics	Timing of environmental impacts captured	Robust causal relationships	Assessment of net-impacts	Consistent micro-macro linkages	Appropriateness of indicators and methods…	Unambiguous and understandable results…
GNB	Propensity score matching	Medium	Medium	High	Medium	Medium	Medium	Medium

The reflection and discussion how different stakeholder priorities fit with high and low scores was conducted in two steps.

Step 1:

The first step was to review the performance assessment and the impact of different stakeholder priorities (weights) only with the evaluation approach tested in the own case study. As not all, or maybe in some cases even the majority of, stakeholders were very familiar with the methodological approaches tested in the case studies, we did not suggest to conduct a detailed validation exercise of the performance assessment at this stage. The purpose of this part of the workshop was rather to reflect how the performance across the criteria fits with the different stakeholder priorities and how their different priorities affect the result of the assessment.

Performance scores shown in the performance matrix (Table 7) of the most (and least) important criteria for the different stakeholders were compared and the differences as well as the potential consequences for the suitability of the tested evaluation approach were discussed.

Step 2:

The second step was to compare the impact of the different stakeholder priorities on the result of the performance assessment of two different approaches (from two case studies – for the example of the German water quality case study we used the structural model approach from the Finnish case study) to be able to highlight that the outcome might be the selection of a different approach.

This required a short explanation of the evaluation approach tested in the other case study explaining the differences in the scores in the performance matrix (both approaches were included in one matrix – two different rows – see Table 8). Then the impacts of the different stakeholder priorities of the different criteria were shown for example stakeholders who would give the highest weights to criteria such as Appropriateness of indicators and Unambiguous and understandable results would select the GNB-DiD approach in Germany while stakeholders who would give the highest weights to criteria such as Timing of environmental impacts captured and Establishment of consistent micro-macro linkages would select the structural model approach.

Table 7 Illustrative example of performance matrix for two case studies / evaluation approaches

Evaluation approach		Quality criteria						
		Responsiveness		Analytical soundness			Measureability	Ease of interpretation
		Judgement (performance) criteria						
Indicator	Method	Compatibility with local environmental … characteristics	Timing of environmental impacts captured	Robust causal relationships	Assessment of net-impacts	Consistent micro-macro linkages	Appropriateness of indicators and methods…	Unambiguous and understandable results…
GNB	Propensity score matching	Medium	Medium	High	Medium	Medium	Medium	Medium
GNB	Biophysical and structural modelling	Medium	Low	High	Medium	Medium	Medium	High

3.2.3 Internal validation of the performance assessment

In addition to the stakeholder-based validation an internal validation approach has been implemented to increase the robustness of the performance assessment as part of the case study testing. A template has been developed which, in addition to defining the impact level for each judgement criteria, required each partner to provide a detailed justification for the performance assessment which relate to:

- the definition of the performance level
- the key features and aspects which led to this level
- the explanation why not a lower or higher performance level has been achieved.

The template was added to the table of the performance assessment shown in Table 2. The initial performance assessment and its justifications and explanations were circulated and reviewed by the other partners acting as referees. The resulting revisions of the performance assessment were systematically recorded and reported using a common reporting template for each judgement criteria.

Reporting template for revisions to the performance assessment (example for the first:

Case study: _____

Approach 1: _____

Performance assessment, what adjustments have been agreed?

Criteria: Compatibility with local environmental characteristics

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Timing of environmental impacts captured

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Establishment of robust causal relationships

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Assessment of net-impacts

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Establishment of consistent micro-macro linkages

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Appropriateness of indicator/method

Change in performance level (none or from to): _____

Justification:

_____

Criteria: Unambiguous and understandable results and policy recommendations

Change in performance level (none or from to): _____

Justification:

_____

The results of the performance assessment are reported in detail for each tested evaluation approach in Deliverable D6.3. A summary of the performance assessment is provided in section 4.2.2.

4 Results of the Assessment

4.1 Cost assessment

The cost of the tested evaluation approaches of the ENVIEVAL case studies were assessed using the cost templates. For each tested approach a cost template for the evaluation cycle was filled in order to cover the whole application process of the evaluation approach. These cost templates are the basis for the cost assessment. It is important to mention that these case studies depict country-specific cases and some of the costs are based on estimations. The cost assessment aims to test a structured approach to assess the cost of an evaluation approach. The conducted analysis should be seen as examples how the collected data can be analysed. The results of the assessment cannot be generalised and easily compared with each other.

4.1.1 Determinants of cost

To assess the cost of an evaluation method it is important to analyse at what steps and activities within the evaluation process decisions are necessary that determine the cost as well as the quality of the outcome of the evaluation. Therefore, it is important to assess the importance of the evaluation steps and the activities in impacting on the overall cost. Further, relevant cost components need to be identified and quantified for these relevant steps of the evaluation process.

4.1.1.1 Activities in the evaluation process

The cost assessment is constructed along the five main phases of the evaluation cycle. These are the evaluation design, the data generation step, the database development and maintenance, the application of the evaluation method and the interpretation of the results. These steps represent the different activities that are conducted in the evaluation process. The information on the cost is collected for the activities within these five evaluation phases according to the steps of the logic model. This enables a detailed attribution of cost to the activities. In this assessment only the cost for the five evaluation phases are compared as this level is sufficient to show the importance of the cost in each evaluation phase.

4.1.1.2 Main cost components

As mentioned above, standard cost components are: a) Labour cost (considering required skills), b) Contracting cost, c) Equipment and other consumables, d) Travel cost, e) Indirect cost and f) Transaction cost. As none of the case studies reported any transaction costs or “other” cost components, these two activities will be excluded from the analysis of the cost. The assessment

focuses on the remaining five main cost components: labour, contracting cost, equipment and consumables, travel cost and indirect cost.

4.1.2 Comparison of cost of the public good case studies

Costs vary strongly between different member states due to differences in the cost levels, e.g. in the case study areas the cost for a staff hour vary which influences the overall cost of an evaluation method. Therefore, the analysis uses the relative cost of the different evaluation steps to ensure comparability and suitability for inclusion in the cost-effectiveness synopsis as well as to account for the different conditions in the case study areas. Out of our 14 public good case studies, 20 evaluation approaches were tested and their costs were reported and feed into this analysis. Four public goods were selected, namely water quality, climate stability, biodiversity (High Nature Value farmland) and landscape to represent the ENVIEVAL public good case studies and are included in the analysis. The case studies were selected as they well represent the varying conditions and evaluation approaches tested in the case studies.

In some of the case studies several methodological approaches were tested (e.g. Water quality in Germany and Greece tested two evaluation approaches, and Landscape in Scotland three). In the assessment only one approach per case study region is included as the cost of the evaluation approaches within one case study region are very similar (in some cases equal).

4.1.2.1 Comparison of different evaluation steps and activities

The share of the cost of the five evaluation phases is compared between the selected case studies to show their importance in varying conditions of different public goods, methods and countries.

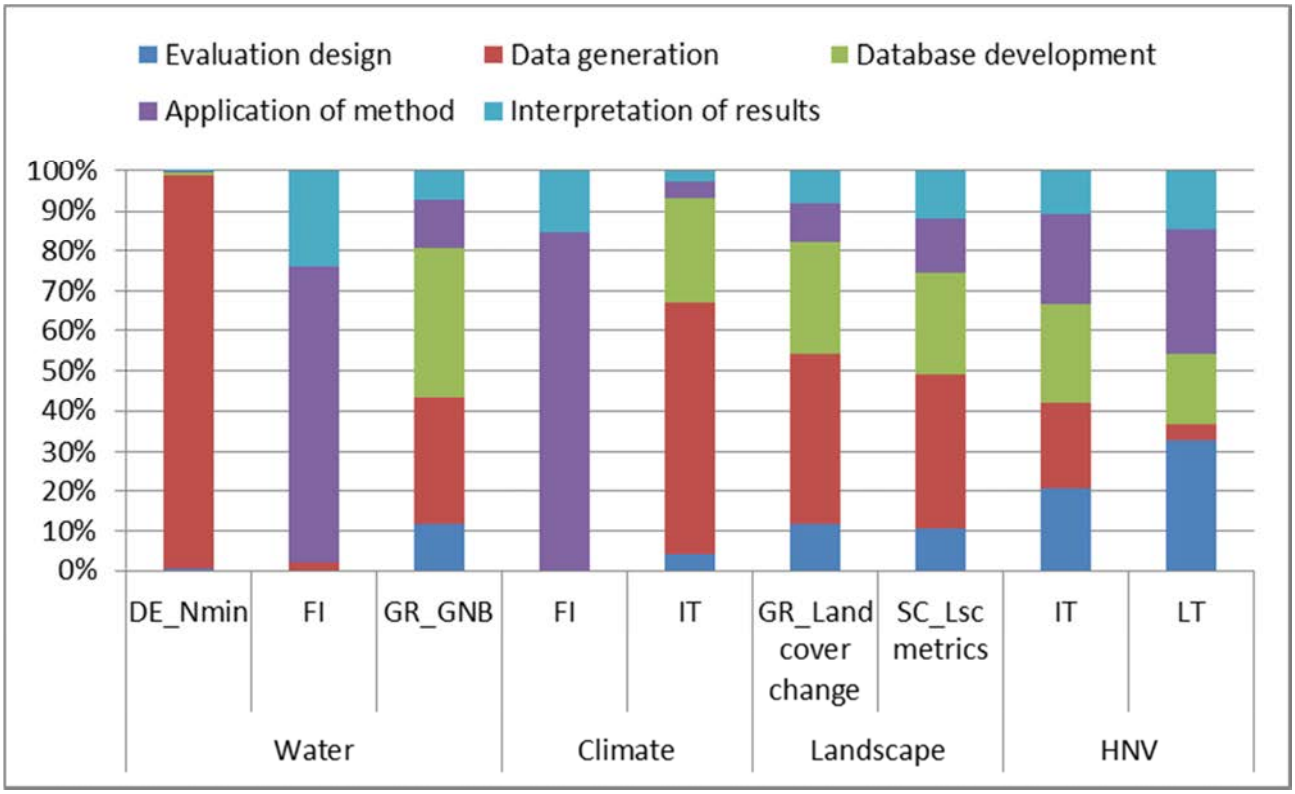

Figure 8 Comparison of the different activities in the evaluation process

It is obvious that costs vary strongly between the different evaluation activities of the ENVIEVAL case studies. This can be traced back to the large diversity of evaluation methodologies and data availabilities in the case study areas. In several case studies (water quality in Germany, climate stability in Italy and the landscape case studies) data generation has the highest share of the total cost. This emphasises the importance of costs related to the use of existing data or additional data collection. This is particularly the case in the German water quality case study, where the collection of monitoring data accounts for 98 % of the cost. Also in the Italian climate case study, the high costs are related to the conduction of a survey by the evaluators. In the Finnish case studies, the application of the method has the largest share with 74% and 85 % of the total cost. Those case studies use existing models, thus the evaluation design, data generation and database development were conducted beforehand, and cost related to these activities are not included in this analysis. The 'interpretation of the results' are the second important phase in these two cases. The other case studies, such as the Greek water quality case study and the HNV case studies, the distribution among the five phases is more even.

To sum up, the importance of the evaluation phases for the cost of the evaluation approaches varies between the case studies and depends on factors such as the use of data and specific conditions of the public goods and the partner countries.

4.1.2.2 Comparison of importance of different cost components

Relevant cost components need to be identified and quantified for the relevant steps of the evaluation process. Standard cost components are: Labour and personnel (considering required skills), contracting, equipment and other consumables, travel cost, and indirect cost. The importance of these cost components is represented in the following figure.

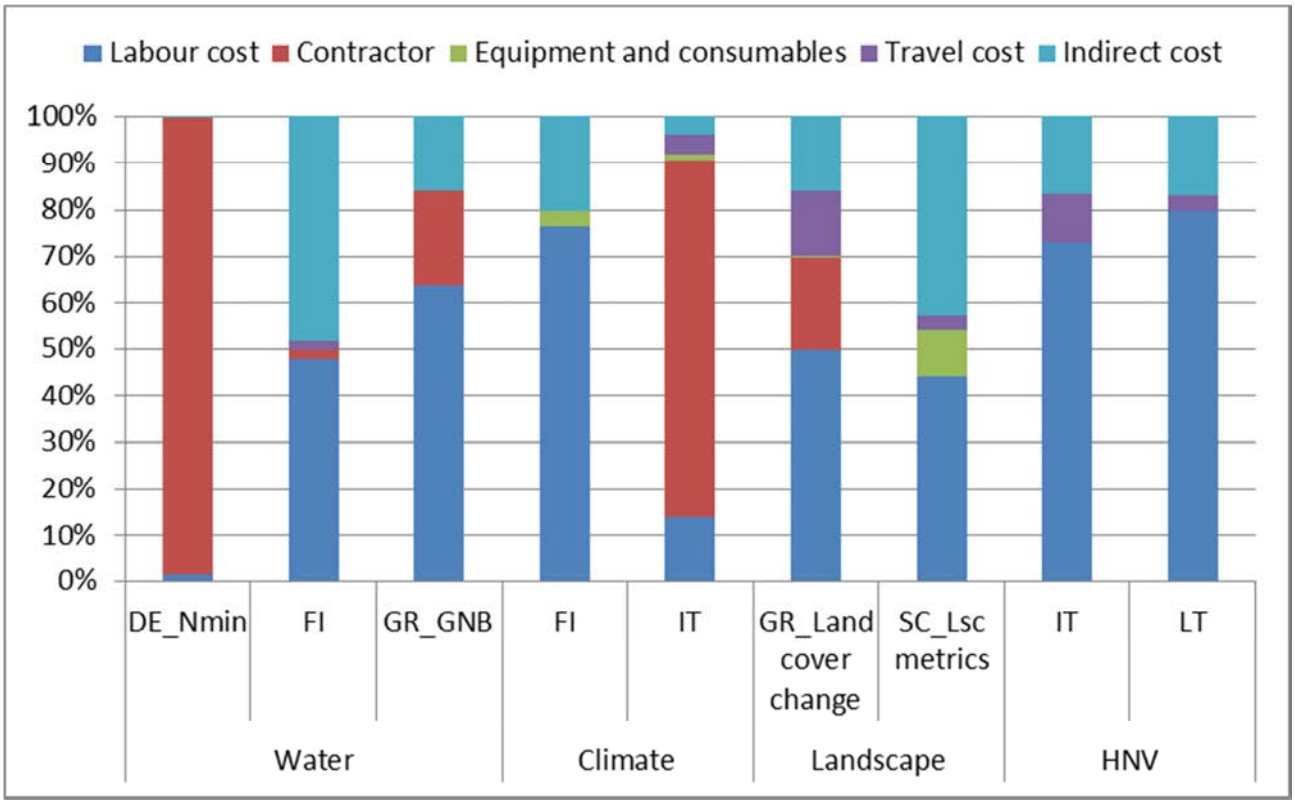

Figure 9 Comparison of importance of the main cost components

It becomes clear that labour cost is the main cost component in nearly all considered public good case studies. Only if contracting is used for the evaluation method is the importance of labour cost lower. This can be explained by the fact that labour cost is included in the contracting cost as it represents the overall cost of a task. Contracting is mainly related to the collection of additional data. In the German Water quality case study and the Italian climate case study, contracting cost have a share of 98 % and 77% respectively. That could mean that for the evaluator costs do only occur for the application of the evaluation method.

Minor cost sources are usually other cost components such as equipment and consumables and travel cost. Indirect cost rates are usually calculated at around 20% of the total cost. When contracting is included, the indirect cost are lower than 20% as the indirect cost are already included in the contract and are not accounted for in the total cost.

It can be concluded that labour and contracting are usually the driver of the overall cost of the evaluation method while other cost components seem to be less important.

4.1.2.3 Integration and importance of monitoring cost

The integration of monitoring cost is an important issue in this assessment. Two cost threads were identified that should be included in the assessment: First, the costs that arise for the evaluator and thus for the application of the evaluation method in itself. This analysis should provide information on the cost of the application of the tested evaluation methods for the evaluation organisations. Second, the overall cost of the evaluation method, including monitoring cost, have to be included. Also when the monitoring data can be accessed for free by the evaluator, the cost need to be attributed to the evaluation method to get an holistic view on the total cost of the evaluation method. If monitoring data is not existent, the cost of data collection has to be attributed to the evaluation method as well. In this section, the importance and differences between the different cost threads are analysed.

Cost that arise for the evaluator and thus for the application of the evaluation method are assessed. This analysis should provide information on the cost of the application of the tested evaluation methods for the evaluation organisations. Cost of data that is not collected by the evaluation entity itself is not taken into account except when costs occur due to the purchase of data. Monitoring data collected by a third party, which could be the monitoring organisation or a contractor, is also not taken into account as the costs do not arise directly for the evaluator. This provides information on the cost of an evaluation method for the evaluator which strongly influences its selection. It is probable that an evaluator chooses evaluation methods which are not associated with high costs or efforts for data collection when he expects to receive appropriate results.

The overall cost of the evaluation method, including monitoring cost, have to be considered. Also when the monitoring data can be accessed for free by the evaluator, the costs need to be attributed to the evaluation method to get a holistic view on its total cost. If monitoring data is not existent, the cost of data collection has to be attributed to the evaluation method as well.

Figure 10 shows the comparison of the two cost threads for the selected case studies of the public goods water quality, climate stability, landscape and biodiversity HNV. The costs for using monitoring data (MO) are opposed with the cost that arises directly for the evaluator when using a certain evaluation method (E). The evaluation costs (E) are expressed as the share of the overall cost of the evaluation approach when monitoring costs are included. Case studies that are not using any monitoring data are excluded from the assessment as no changes are expected. This refers to

the Lithuanian HNV case study and the Finnish case studies for Water quality and climate stability.

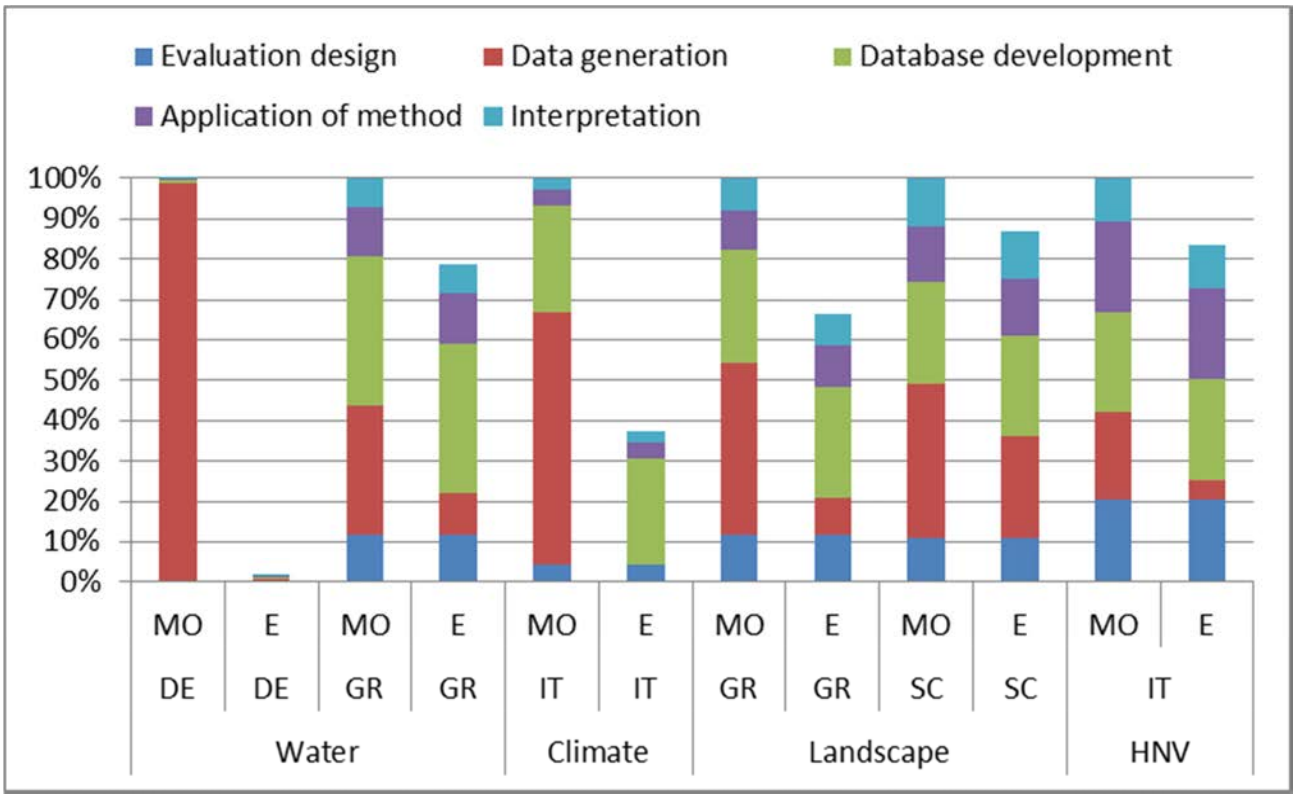

Figure 10 Comparison of evaluation cost and monitoring cost

The comparison of the two cost strains identifies the importance of data generation as this is the step where monitoring data is generated and the costs occur. In the data generation step, situations differ because of the different circumstances in the case studies. In the German water quality case study as well as the Italian climate case study, the total cost of the evaluation approach decreases strongly when deducting the monitoring cost. Particularly in the German case, the cost reduction is high as only 2 % of the total cost remains when monitoring cost is excluded. Further it can be seen that for these two case studies after the deduction of the monitoring cost, only few or no costs occur in the data generation step. In other case studies, the share of the cost of the data generation step is reduced when monitoring costs are deducted. However cost for the use of existing data still occurs.

It is obvious that costs vary strongly between the different evaluation activities of the ENVIEVAL case studies. This can be traced back to the large diversity of evaluation methodologies and data availabilities in the case study areas. However, it is obvious that when data collection by evaluators or monitoring organisations is necessary it is often the main cost source. Compared to the additional data collection, the use of existing data usually seems to be associated with lower

costs. Reasons for the lower cost could be on the one hand that suitable data for the evaluation methodologies do not exist or that the purchase is associated with high cost that cannot be paid by the evaluator. On the other hand, in some member states (e.g. in Germany) the data is provided free of charge to the evaluators which leads to the low cost for the use of existing data sets. A reason for this is that the monitoring is performed also for other purposes, e.g. for general statistics, technical advice, enforcement of mandatory standards of good farming practice, or monitoring of local conditions. In these cases no additional cost is attributed to the use of existing data for evaluation.

4.1.2.4 Comparison of relative cost of different evaluation method that refer to similar public goods

For the comparison of different approaches of the same public good case study, the water quality and landscape case studies were selected as both have, with five different methodological approaches, the highest variety of approaches for one public good. Therefore, the differences of the cost composition are analysed more in detail. Please note that for the German case studies, monitoring costs were excluded from the analysis as they were provided free of charge to the evaluators. As the other approaches also do not include monitoring cost, these costs were excluded from the analysis to be able to compare the cost for the evaluator. The following figure includes the relative cost of the water quality evaluation approaches.

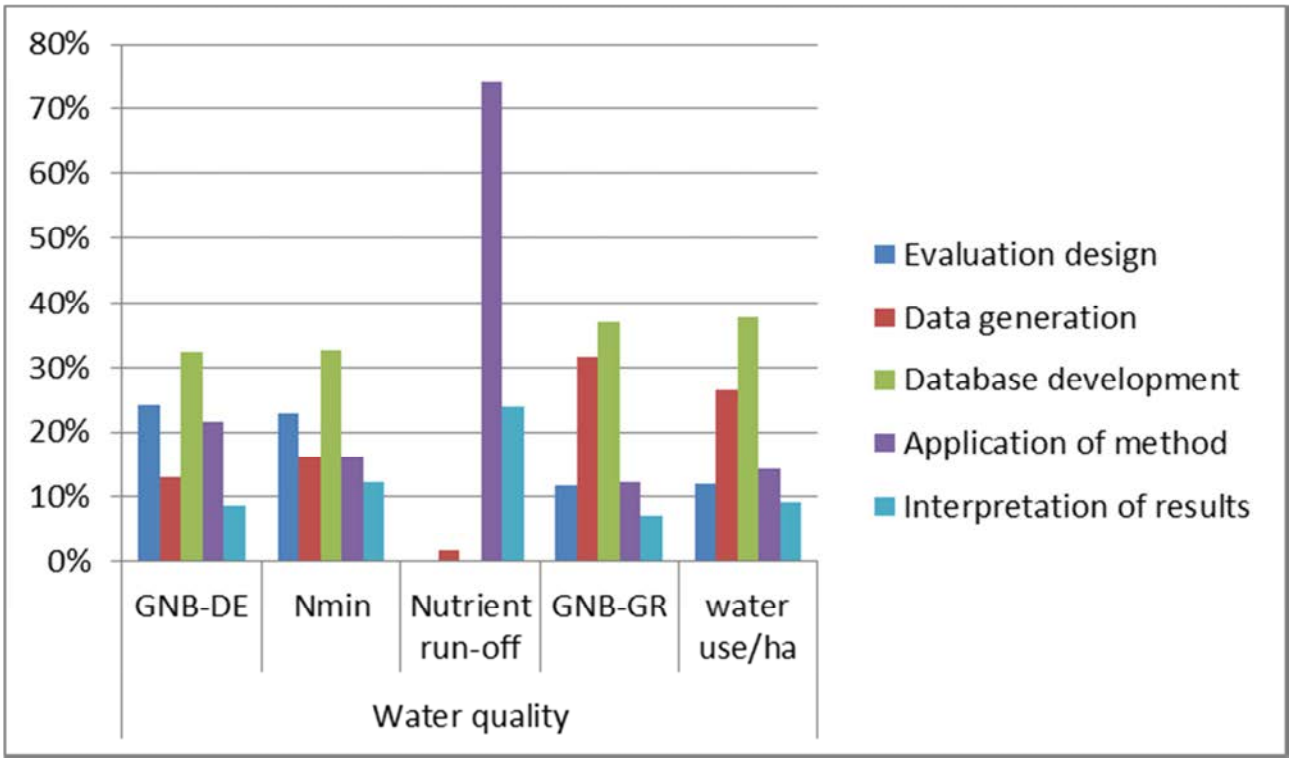

Figure 11 Comparison of different evaluation approaches of the water quality case studies

Water quality case studies are conducted in three partner countries. Five different indicator/method combinations are used. Looking at the approaches, it can be said that the composition of the cost varies between the case studies while different approaches in the same country seem to be very similar (GNB-DE and Nmin in Germany, and GNB-GR and water use/ha in Greece). This could be explained by the use of similar infrastructures and maybe that the same staff members conducted the analysis. Further, the Greek and the German case study used the CMEF indicator gross nutrient balance (GNB) so the application of the same indicator in two countries can be explored. However, the methodologies used for the evaluation differ which inhibits the comparison. It can be said that the cost composition of two approaches of the same partner country are more similar than the two approaches using the same indicator.

The Finnish case study, as mentioned before, has the largest share of the cost for the application of the method followed by the interpretation of the results. As the case study used a modelling approach, it can be concluded that the required resources of this modelling approach differ strongly from other approaches which are using monitoring data and/or biophysical models.

The landscape case studies compose of two approaches for the Greek case study (land cover change and visual amenity) and three approaches for the Scottish case study (landscape metrics, Natura 2000 and visibility of change). The approaches are compared in the figure below.

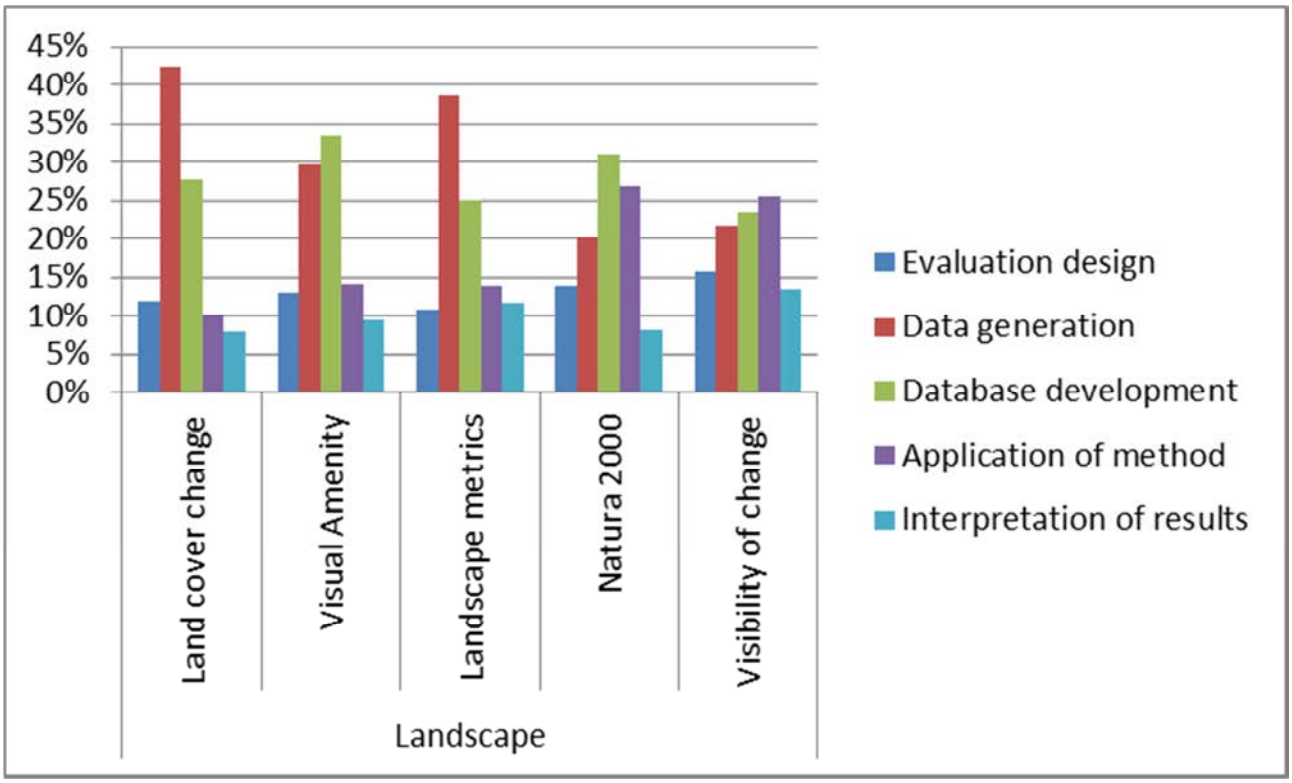

Figure 12 Comparison of different evaluation approaches of the landscape case studies

The costs of the evaluation approaches do not differ as much as the approaches of the water quality case studies. Data generation and database development have usually the largest share of the cost. The Natura 2000 and the Visibility of change approach of the Scottish case study show the highest share of cost for the application of the method. Again it can be concluded that approaches from the same country have very similar compositions.

4.1.3 Implications for evaluations

The report provides an overview of the cost of the ENVIEVAL public good case studies and should serve to establish a framework for the assessment of the cost of monitoring and evaluation of RD programmes. The data enables the comparison of the cost for the different evaluation steps and the main cost components. As cost vary strongly between the different case studies due to different conditions in the member states, different approaches and data requirements only the comparison of the relative cost is reasonable.

The results show, that although cost vary strongly between the case studies some trends are visible:

- In many cases data generation has the highest share of cost. This can be traced back to the high cost for additional data collection. When modelling approaches are used (e.g. Finnish case studies) the application of the evaluation method is very time intensive and has a high demand for cost.
- Labour costs seem to be the main cost component and is usually the driver of the overall cost. If contracting costs are involved in the evaluation approach, this is the main cost source. Labour costs are included in the contracting cost.
- The cost for the evaluator and the managing authority can be very different depending on the availability of monitoring data. Therefore, it is important to distinguish between the two cost threads.
- The costs of the application of different evaluation approaches in one country seem to be very similar. Cost variations are higher between similar approaches in different countries. This could mean that the transferability to other countries is more limited as each country experiences unique conditions.

The cost templates could help evaluators to plan and control evaluation cost in a structured way and to identify the main drivers of cost. The comparison of costs of evaluation approaches remains challenging although the detailed assessment of cost helps to show the drivers of cost for each

evaluation approach. Comparability is further limited due to different conditions in the partner countries (e.g. different data access and expertise for statistical analysis) and evaluation agencies. This shows that the mere comparison of cost of evaluation approaches is not sufficient. It is also important to consider the effectiveness of the approaches in order to get a holistic valuation of the cost-effectiveness of evaluation approaches.

4.2 Effectiveness assessment

The aim of the effectiveness assessment is to apply an approach to assess the performance of the tested evaluation approaches in addressing the main evaluation challenges in the context of different stakeholder priorities. In this section we first present the results of the definition of different stakeholder weights (priorities) for the various judgement criteria used in the performance assessment followed by a summary of the performance assessment of the tested evaluation approaches highlighting how different stakeholder priorities affect the interpretation of the results and ultimately the selection of the approach for environmental impact evaluations of RDPs. This section also provides the basis for the assessment of different scenarios for improving the strategic design of environmental monitoring programmes in the cost-effectiveness synopsis in section 5.2.

4.2.1 Defining weights for the performance (judgement criteria): Results of the participatory stakeholder workshops

In order to generate weights for the different judgement (performance) criteria, the stakeholders validated and assessed the importance of the criteria in participatory exercises. Table 9 shows the average stakeholder priorities across all stakeholder types in the partner countries based on the allocation of the priorities (15 dots of the seven criteria – see section 3.2.2 for more details) in the workshops.

On average the criteria ‘Appropriateness of indicators’, ‘Establishment of robust causal relationships’ and ‘Assessment of net-impacts’ had the highest stakeholder priorities, while the ‘Establishment of consistent micro-macro linkages’ was generally of lesser importance for the stakeholders. There are, however, significant differences between the stakeholder priorities in the partner countries. The largest variation in priorities between the different criteria was expressed in Hungary with a very low importance for the ‘Compatibility with local environmental and farm structural characteristics’ and a particular high importance for ‘Timing of environmental impacts captured’. In contrast, stakeholders in Scotland have allocated relative equal priorities across the criteria between with most of the scores ranging between 2.0 and 2.4. In the context of specific

criteria, stakeholders in Germany have noticeably given a lower priority to the Appropriateness of indicators than stakeholders in the other partner countries.

Table 9 Average stakeholder priorities of effectiveness criteria in partner countries

Judgement criteria	Hungary	Italy	Germany	Scotland	Lithuania	Finland	Greece	Average all
No of Stakeholders	11	8	6	6	12	6	7	56
Compatibility with local environmental and farm structural characteristics	0,7	2,4	2,2	2,2	1,7	2,5	2,0	2,0
Timing of environmental impacts captured	3,8	1,9	1,7	2,0	1,7	2,0	1,7	2,1
Establishment of robust causal relationships	3,1	2,3	2,8	1,6	2,9	1,8	2,3	2,4
Assessment of net-impacts	1,6	1,8	3,0	2,2	2,5	2,5	2,7	2,3
Establishment of consistent micro-macro linkages	1,3	1,9	1,7	2,2	1,3	1,7	1,6	1,7
Appropriateness of indicators and methods to capture complexity of environmental relationships	2,4	2,6	1,5	2,4	3,0	2,5	2,9	2,5
Unambiguous and understandable results and policy recommendations	2,1	2,3	2,2	2,4	2,2	2,0	1,9	2,1
Test (should add up to 15)	15	15	15	15	15	15	15	

However, looking at the differences between partner countries in more detail in Table 10 synthesises the average priorities of evaluators and representatives from managing authorities and monitoring organisations in the partner countries.

Table 10 Synthesis of average priorities of evaluators and MAs / MOs

Stakeholders	Evaluators			MA / MO		
Criteria /average no. dots	< 2	2	> 2	< 2	2	> 2
Compatibility with local environmental and farm structural characteristics	DE, HU	GR, LT	FI, SCO	HU, LT	FI, SCO	DE, *GR*
Timing of environmental impacts captured	DE. LT	FI, GR, SCO	*HU*	GR	DE, FI, LT, SCO	HU
Establishment of robust causal relationships	FI	GR, SCO	DE, HU, LT	-	DE, FI, GR, SCO	LT, *HU*
Assessment of net-impacts	-	GR, HU, SCO	DE, FI, LT	-	FI, GR, HU, LT, SCO	DE
Establishment of consistent micro-macro linkages	-	DE, FI, GR, HU, LT, SCO	-	DE, FI, *GR*, HU, LT	-	SCO
Appropriateness of indicators and methods to capture complexity of environmental relationships	DE, HU	FI, GR, SCO	LT	-	DE	FI, *GR*, HU, LT, *SCO*
Unambiguous and understandable results and policy recommendations	-	FI, GR, HU, LT	DE, SCO	DE	FI, SCO	GR, HU, LT
Italics - indicates a allocation of dots higher than 3 or 0						

Table 10 shows that the criteria 'Unambiguous and understandable results and policy recommendations' is of high priority across stakeholder types and countries. As also shown in Table 9, the criteria 'Establishing consistent micro-macro linkages' is of relatively low importance, in particular for those participating stakeholder which have represented managing authorities and monitoring organisations.

Looking at the differences between evaluators and representatives from managing authorities & monitoring organisations, the results suggest that evaluators give more importance to criteria linked to the analytical soundness such as 'Establishment of robust causal relationships' and 'Assessment of net-impacts' while representatives from managing authorities & monitoring organisations give higher priorities to the 'Appropriateness of indicators'.

An important aspect to be taken into account in the assessment of the differences in emerging stakeholder priorities is that the expressed priorities are dependent on individual characteristics such as preferences, methodological competences and resources and can be influenced by different interpretations of the (scope and definition of the) criteria. While care has been taken to use the same detailed explanations in each workshop, some differences in the interpretation of the criteria have emerged. For example, the low priority of Finnish evaluators for the criteria 'Establishment of robust causal relationships' is directly linked to their understanding that this criteria is included in the choice of an appropriate indicator, which has received a higher priority. Also, the high priorities for the criteria 'Unambiguous and understandable results and policy recommendations' were only defined after clarification that this criteria also refers to clear and transparent policy recommendation, which are strictly based on the results of the evaluations.

The discussions at the workshops highlighted the complexity of trying to achieve a better understanding of the relationships and linkages between criteria and the impact on the performance (effectiveness) assessment of the evaluation approaches. A possible hierarchical order of criteria (e.g. timing and local conditions as precondition for establishing causal relationships) could be highlighted in a conceptual model (hierarchy) of key effectiveness aspects to be considered in the design and selection of evaluation approaches in the methodological handbook.

Despite some differences in the interpretation of the criteria, the results of the participatory workshops could be used to derive a set of different weights of the judgement criteria to highlight the possible impact of stakeholder priorities on the selection of the most suitable evaluation approach. Table 11 derives the average weights defined by evaluators and managing authorities &

monitoring organisations in the partner countries taking into account the allocation of dots and the results of the pair-wise comparisons (see section 3.2.2 for details).

Table 11 Definition of weights for performance assessment

		Judgement (performance) criteria						
Country	Stakeholder	Compatibility with local environmental … characteristics	Timing of environmental impacts captured	Robust causal relationships	Assessment of net-impacts	Consistent micro-macro linkages	Appropriateness of indicators and methods	Unambiguous and understandable results…
LT	E	High	Medium	Very high	Very high	High	Very high	High
	MA / MO	Medium	High	Very high	High	Medium	Very high	Very high
FI	E	Very high	High	Medium	Very high	High	High	High
	MA / MO	High	High	High	High	Medium	Very high	High
DE	E	Medium	Medium	Very high	Very high	High	Medium	Very high
	MA / MO	Medium	Medium	High	Very high	Medium	Medium	High
GR	E	High	High	High	High	High	High	High
	MA / MO	Very high	Medium	High	High	Medium	Very high	Very high
SCO	E	Very high	High	High	High	High	High	Very high
	MA / MO	High	High	High	High	Very high	Very high	High
HU	E	Medium	High	Very high	High	High	Medium	High
	MA / MO	Medium	High	High	High	Medium	Very high	Very high
IT	E	Very high	High	High	Medium	High	High	Very high
	MA / MO	High	High	Very high	High	Medium	Very high	High
All	E	High	High	High	High	High	High	High
All	MA / MO	High	High	High	High	Medium	Very high	High
Average		High	High	High	High	Medium	High	High
		1,95	2,11	2,40	2,34	1,49	2,46	2,14

A different set of weights, differentiating between three weight categories medium, high and very high, was defined for evaluators and managing authorities & monitoring organisations in each partner country. The definition of the weight levels are based on the following classification:

- medium: < 2 dots
- high: 2 dots (average)
- very high: > 2 dots

The defined set of weights reflects the identified differences in stakeholder priorities shown in Table 9 and Table 10. The different set of weights enabled a comparative assessment of the performance of the tested evaluation approaches in comparison to equal weights across all judgement criteria. In other words, the weights could be used to review to what extent the tested evaluation approaches were able to address the main evaluation challenges which were of highest priority for the different types of stakeholders in each partner country and on average across all partner countries. In addition, the different weights and identified stakeholder priorities were used to validate and reflect on the stakeholder relevance of the potential contributions of the monitoring data scenarios assessed in the cost-effectiveness synopsis. The assessment provides a better

understanding of the differences between stakeholder priorities and their implications for evaluations.

4.2.2 Performance assessment of the tested evaluation approaches

Detailed assessments of the performance of the tested evaluation approaches using the framework developed in section 3.2 have been reported in the case study summary reports in Deliverable D6.3. Here only a short summary of the performance matrix of the tested evaluation approaches is provided. The main purpose of this section is to explain examples of comparative assessments between different evaluation approaches and to highlight how different stakeholder priorities affect the interpretation of the results and ultimately selection of the approach for environmental impact evaluations of RDPs.

Overview of the performance assessment

Table 12 shows the performance matrix for the evaluation approaches tested in the environmental public good case studies. The tested evaluation approaches were most successful in achieving high performance or impact levels with respect to the criteria 'Establishment of robust causal relationships' and 'Unambiguous and understandable results and policy recommendations'. 14 of the listed twenty tested evaluation approaches achieved a high impact level for those two criteria followed by the criteria 'Appropriateness of indicators and methods' to capture complexity of environmental relationships with 13 high impact levels assigned (Table 13).

The high performance levels for the 'Establishment of causal relationships' and the 'Appropriateness of indicators and methods' to capture the complexity of environmental relationships highlights the emphasis of the public good case studies on contributions of additional (non-CMES) indicators tested to address indicator gaps and contributions of advanced modelling approaches tested at micro and macro level for dealing with the complexity of public goods (see also the discussion section of Deliverable D6.3).

At the opposite end, only 3, respectively 4, tested evaluation approaches achieved a high performance level for the criteria 'Establishment of consistent micro-macro linkages' and 'Assessment of net-impacts', which reflects the severity of the methodological challenges underlying those two criteria as well as the large data requirements of evaluation approaches able to address these challenges.

Table 12 Performance matrix with impact levels: Overview of the tested evaluation approaches

Public good case study		Evaluation approach		Quality criteria						
				Responsiveness		Analytical soundness			Measureability	Ease of interpretation
				Judgement (performance) criteria						
Public good	Country	Indicator	Method	Compatibility with local environmental … characteristics	Timing of environmental impacts captured	Robust causal relationships	Assessment of net-impacts	Consistent micro-macro linkages	Appropriateness of indicators and methods…	Unambiguous and understandable results…
Biodiversiry	HU	Number of farmland bird species (NBS)	Spatial analyses of survey spots	Medium	High	High	High	Low	High	High
	HU	Farmland Bird Index	Spatial analyses of quadrats	High	High	High	Medium	Low	High	Medium
	LT	Singing males of corncrake	Multiple regression analysis & upscaling	High	Low	High	Medium	Medium	Low	High
	LT	White stork breeding density	Multiple regression analysis & upscaling	High	Low	Medium	High	Medium	High	High
Climate stability	FI	CO2 equivalent measures	Partial equilibrium model	Low	High	High	High	Medium	High	High
	IT	GHG balance	Carbon footprint	Medium	High	Low	Medium	Medium	High	High
HNV	IT	% of HNV farmland, HNV score	Multicriteria analysis & upscaling	High	Medium	Medium	Low	Medium	High	Medium
	LT	Changes in diversity of ecotones	Spatial analysis	High	High	Medium	Medium	High	High	High
Landscape	GR	Land cover change	Spatial analysis (DiD)	High	High	Medium	Low	Medium	Medium	High
	GR	Visual amenity	Spatial analysis (DiD)	High	High	High	Low	High	High	High
	SCO	Patchshape index	Landscape metrics	High	Medium	Medium	Medium	High	High	Medium
	SCO	Visibility of change	Spatial analysis	Medium	Medium	High	Medium	Low	High	High
Soil quality	HU	Soil organic carbon content	CLUE model	Medium	High	High	High	Medium	Low	Medium
	SCO	Soil organic carbon content	INVEST model	Medium	Medium	High	Low	Medium	Low	Low
	SCO	Annual average soil loss and sediment retention	USLE model	High	Medium	High	Low	Medium	Medium	High
Water quality	DE	Nmin	Pairwise comparisons & regression analysis	Medium	Medium	High	Medium	Medium	High	High
	DE	GNB	Propensity score matching	Medium	Medium	High	Medium	Medium	Medium	Medium
	FI	GNB	Biophysical and structural modelling	Medium	Low	High	Medium	Medium	Medium	High
	GR	GNB	Biophysical modelling	High	Low	High	Medium	Low	High	High
	GR	Water use/ha	Biophysical modelling	High	Low	High	Medium	Low	High	High

Table 13 Number of impact levels achieved for the different criteria

Judgement (performance) criteria	Performance assessment: Frequency of impact levels		
	High	Medium	Low
Compatibility with local environmental and farm structural characteristics	11	8	1
Timing of environmental impacts captured	8	7	5
Establishment of robust causal relationships	14	5	1
Assessment of net-impacts	4	11	5
Establishment of consistent micro-macro linkages	3	12	5
Appropriateness of indicators and methods to capture complexity of environmental relationships	13	5	2
Unambiguous and understandable results	14	5	1

However, the performance assessments do not reflect the full potential of the tested approaches as a number of constraining factors limited the performance in the case study testing. Issues in relation to particular skills required applying certain evaluation approaches and administrative and institutional issues (e.g. length of evaluation contracts providing sufficient time and resources to apply more advanced and complex evaluation approaches) constraint the effectiveness of evaluation. But the performance assessment of the evaluation approaches carried out in the case studies highlights data issues as the single most important factor influencing the effectiveness of the evaluation approaches. The results of the case studies indicate that the cost-effectiveness of monitoring programmes and environmental evaluations can be improved through strategic sampling of environmental monitoring data. More targeted environmental monitoring programmes would facilitate a more robust quantification of deadweight effects and causal relationships and other intervening factors. Hence, the synopsis in section 5.2 will focus on assessing the cost-effectiveness implications of monitoring data scenarios.

Comparative assessment in the context of different stakeholder priorities

The first step in the comparative assessment is the qualitative sign analysis of the performance matrices done by each partner. A qualitative sign analysis is a simple synthesis tool which is based on the assumption that qualitative data cannot be added up, but that the frequency of occurrence of an impact level over a range of criteria can be numerically treated (Nijkamp and Blaas, 1994).

The number of high, medium and low scores of each methodological (evaluation) approach of the performance assessment in the performance matrix is synthesized in a frequency table. Applying different weights to the judgement criteria of the performance assessment the frequency of different scores can then separately be summarised for each of three weight categories (i.e. for the criteria given a very high priority by the stakeholders, for the criteria with high priority and criteria with medium priority). The starting point, however, is equal weights across all criteria and an overview table summarising the frequencies for the whole performance matrix of all tested evaluation approaches (Table 14).

Table 14 Frequency table for all tested evaluation approaches (equal weights for criteria)

Public good case study		Evaluation approach		Number of judgement criteria per impact level		
Public good	Country	Indicator	Method	High	Medium	Low
Biodiversity	HU	Number of farmland bird species (NBS)	Spatial analyses of survey spots	5	1	1
	HU	Farmland Bird Index	Spatial analyses of quadrats	4	2	1
	LT	Singing males of corncrake	Multiple regression analysis & upscaling	3	2	2
	LT	White stork breeding density	Multiple regression analysis & upscaling	4	2	1
Climate stability	FI	CO2 equivalent measures	Partial equilibrium model	5	1	1
	IT	GHG balance	Carbon footprint	3	3	1
HNV	IT	% of HNV farmland, HNV score	Multicriteria analysis & upscaling	2	4	1
	LT	Changes in diversity of ecotones	Spatial analysis	5	2	0
Landscape	GR	Land cover change	Spatial analysis (DiD)	3	3	1
	GR	Visual amenity	Spatial analysis (DiD)	6	0	1
	SCO	Patchshape index	Landscape metrics	3	4	0
	SCO	Visibility of change	Spatial analysis	3	3	1
Soil quality	HU	Soil organic carbon content	CLUE model	3	3	1
	SCO	Soil organic carbon content	INVEST model	1	3	3
	SCO	Annual average soil loss and sediment retention	USLE model	3	3	1
Water quality	DE	Nmin	Pairwise comparisons & regression analysis	3	4	0
	DE	GNB	Propensity score matching	1	6	0
	FI	GNB	Biophysical and structural modelling	2	4	1
	GR	GNB	Biophysical modelling	4	1	2
	GR	Water use/ha	Biophysical modelling	4	2	1

However, the comparability of all tested approaches is limited due to a different emphasis on particular evaluation challenges, micro and macro levels and certain evaluation approaches can only be apply in the context a specific public good. This implies that only a small number of approaches across the same or similar public goods can be compared. A small number of comparable approaches do not merit the application of a more in-depth outranking and trade-off analysis, but a qualitative sign analysis shall be used to highlight how different stakeholder priorities affect the interpretation of the results and ultimately selection of the approach for environmental impact evaluations of RDPs.

An example of evaluation approaches tested in the landscape case studies is used to highlight the impact of different stakeholder priorities on the selection of evaluation approaches[1]. Table 15 shows the frequency table for those two evaluation approaches assuming equal weights across all seven judgement criteria.

Table 15 Frequency table for the example (equal weights assumed)

Public good case study		Evaluation approach		Number of judgement criteria per impact level		
Public good	Country	Indicator	Method	High	Medium	Low
Landscape	SCO	Patchshape index	Landscape metrics	3	4	0
	SCO	Visibility of change	Spatial analysis	3	3	1

Assuming equal weights across the judgement criteria, the results of the performance assessment would suggest a selection of the first approach (based on the slightly better performance) using a patchshape index indicator and landscape metrics as a method.

If one now takes into consideration different stakeholder priorities the decision the interpretation of the performance assessment and the decision which approach to apply might differ. For example let's assume different sets of weights reflecting the stakeholder priorities of the managing authorities and evaluators in Italy as well as of the evaluators in Finland. As shown in Table 11 the representative from the managing authorities in Italy gave a very high priority (weight) to the criteria Establishing robust causal relationships and Appropriateness of indicators and methods to capture complexity of environmental relationships, while the evaluators gave the criteria Compatibility with local environmental and farm structural characteristics and Unambiguous and

[1] Two evaluation approaches tested in the same case study have been selected as example to minimise potential distortions in the performance assessment due to different data availabilities. The example, however, is for illustrative purpose to show the implications of different stakeholder priorities.

understandable results a very high priority. Evaluators in Finland saw the criteria Compatibility with local environmental and farm structural characteristics and Assessment of net-impacts as the most important ones to consider. The stakeholders expect from the evaluation approach a high (or best possible) performance level for those criteria which they have given the highest priority. Table 16 now presents the frequency table for the criteria with the highest stakeholder priority (weight).

Table 16 Frequency table for the highest weight category

Stakeholder group		Managing Authorities Italy			Evaluators Italy			Evaluators Finland		
Weight level		Very high			Very high			Very high		
Performance of evalaution approaches		Impact levels			Impact levels			Impact levels		
Evaluation approaches		High	Medium	Low	High	Medium	Low	High	Medium	Low
Patchshape index	Landscape metrics	1	1	0	1	1	0	1	1	0
Visibility of change	Spatial analysis	2	0	0	1	1	0	0	2	0

Assuming stakeholder priorities as indicated by the representatives from the managing authority in Italy, the second approach can be expected to deliver a higher performance (effectiveness) for the criteria with the highest priority, as both criteria have a high performance or impact level. Stakeholder priorities as indicated by the Finnish evaluators would result in the selection of the first approach, as for the situation with equal weights. Further examination requires the set of stakeholder priorities of the evaluators in Italy. Here, both evaluation approaches have the same performance assessment for the criteria with the highest priority and a decision on which evaluation approach can be expected to deliver a performance more in line with the particular set of stakeholder priorities. Table 17 provides frequency tables for all three weight classes and adds a comparison of the performance for the criteria with a high and medium priority.

Table 17 Frequency table for all three weight categories

Evaluators Italy										
Weight level		Very high			High			Medium		
Performance of evalaution approaches		Impact levels			Impact levels			Impact levels		
Evaluation approaches		High	Medium	Low	High	Medium	Low	High	Medium	Low
Patchshape index	Landscape metrics	1	1	0	2	2	0	0	1	0
Visibility of change	Spatial analysis	1	1	0	2	1	1	0	1	0

It becomes now apparent that under the stakeholder priorities indicated by the evaluators in Italy, the first approach would be more suitable as the performance assessment shows a better result for those criteria with a high priority (middle weight category).

The qualitative sign analysis is an easily applicable multi-criteria assessment method for scoping the cost-effectiveness of alternative evaluation approaches. The qualitative sign analysis can be effective, but it may also lead to the use of unjustified assumptions, causing incorrect ranking of options. In analytically more sophisticated effectiveness assessments, the information in the basic matrix is usually converted into consistent and comparative scores and trade-off matrices are constructed and analysed. This implies that trade-offs between different criteria are acceptable, so that good performance of an evaluation approach on one criterion can in principle compensate for weaker performance on another. But this would require a higher number of comparable evaluation approaches.

The key message of the comparative assessment is that the results of the effectiveness or performance assessment can be differently interpreted depending on the set of priorities attached to the judgement criteria and the final decision which evaluation approach to select often depends on the particular priorities of the stakeholders. The final selection revolves around an inspection of the performance assessment considering the relative costs of the different approaches as well as specific circumstances, preferences and abilities of the end-user (stakeholder). It is however important that a consistent framework is used with clearly defined criteria and performance or impact levels is used.

The identification of stakeholder priorities and their different weights for judgement or effectiveness criteria of evaluation approaches is important for an ex-ante assessment of the potential contributions of possible approaches to select for the evaluation of environmental of RDPs. The development of the conceptual framework with a set of quality and judgement (performance) criteria as well as performance levels is the basis for a robust and sound assessment of the effectiveness of evaluation approaches. The framework developed in the ENVIEVAL project has attempted to fill the gap of a lacking framework and provides a starting point for further improvements and testing of effectiveness assessments of environmental evaluations of RDPs. Crucial in this context is to ensure a good understanding of the linkages and relationships between the different criteria.

The definition of different weights for judgement or effectiveness criteria is also important for an ex-post assessment of the performance (and thus the robustness and quality of the evaluation

results) of the evaluation approaches in the context of different stakeholder priorities. Such an assessment can help to answer to what extent the applied evaluation approaches have delivered the required results, addressed existing evaluation challenges and help to identify the need for further improvements in both the data infrastructure and methodological development.

5 Cost-effectiveness Synopsis

5.1 Main decisions in the evaluation process and their cost

5.1.1 Overview evaluation process

In each step of the evaluation cycle, evaluators have to make decisions that have direct impacts on the cost of the evaluation approach as well as on the effectiveness. In this section, the type of decisions that have to be taken in the evaluation exercise and their effects on the cost-effectiveness of evaluation approaches are analysed. Linkages to the logic model framework that was developed in the ENVIEVAL project are highlighted.

5.1.2 Integration of cost-effectiveness aspects in the methodological framework (logic models)

The assessment of the cost of the evaluation approaches in chapter 3 was developed along the phases of the evaluation process. Each phase consists of several steps[2] where decisions on the development and implementation of the evaluation exercise have to be taken. The following figure shows the five phases of the evaluation cycle and the related steps that influence the evaluation design and thus the cost and effectiveness of the approach.

[2] The steps in Figure 13 reflect the different steps of the logic model based methodological framework developed in the ENVIEVAL project (for more details see Artell et al. (2015), Povellato et al. (2015) and Aalders et al. (2015).

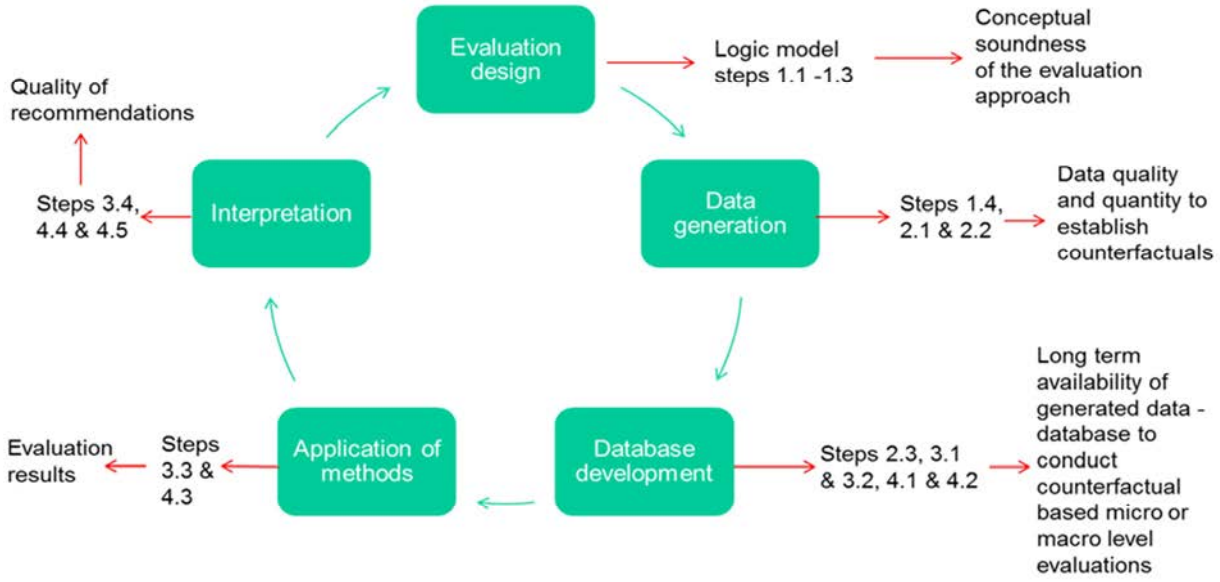

Figure 13 Evaluation cycle, logic model steps and key aspects influencing cost-effectiveness of evaluations

In the first evaluation phase, the evaluation design, several decisions have to be taken, which set the basis for the evaluation approach. The application of the CMES intervention logic has to be applied to the evaluation design (Step 1.1). Additional environmental indicators have to be selected if necessary (Step 1.2). This can be associated with high cost as there is an additional work load for the application of the CMES. However, the selection of suitable additional indicators could increase the effectiveness of the evaluation exercise and might be beneficial. Further, data requirements and available data need to be reviewed for the CMES and additional environmental indicators. These activities are crucial for the successful application of the statistical analysis of counterfactuals and have a strong impact on the effectiveness. High cost savings are possible when existing data sources could be discovered and accessed as data collection is usually expensive. Conceptual decisions have further to be drawn on the selection of the unit of analysis (Step 1.3).

It can be concluded that decisions at the end of the first evaluation phase are mainly associated with increased labour cost as more time is spent on the development of the evaluation approach. However, importantly, these decisions influence the effectiveness of the evaluation approach at all of its other stages. The right decisions at the beginning of the evaluation process are essential for the successful application of the evaluation method and thus merit the higher cost.

The second evaluation phase is associated with data generation activities and includes tasks related to the use of existing data sources and the collection of additional primary data, if necessary. Data is assessed to enable statistical analysis with counterfactual design at a micro and macro level to

enable net-impact assessments. Data availability for counterfactuals needs to be checked (Step 2.1) as well as the possibilities to construct robust counterfactuals with or without comparison groups with the existing data (Step 2.2). If additional primary data collection is conducted, this evaluation step can be at a high cost. The mode of data collection and the sampling strategy have a high impact on the effectiveness of the evaluation as this provides the basis for a sound statistical analysis. The use of existing data sources is usually associated with lower cost as most evaluators have access to a variety of free data sources. Monitoring data are often not directly targeted for use for evaluation purposes and often does not meet the needs of evaluation. This has a strong negative impact on the effectiveness of the evaluation as the results are not robust or the statistical analysis does not cover all aspects of rural development impacts. Thus, increased efforts in planning and design of data collection are worth the improved sampling or coverage of rural development impacts despite the higher labour costs.

In the third phase of the database development and maintenance, important decisions have to be taken which influence the cost-effectiveness of the evaluation approach. The evaluation option for counterfactual based analysis (Evaluation Options without Comparison Groups, Qualitative and Naïve Quantitative Evaluation Options or Statistics-based Evaluation Options) depending on the existing data availability is selected (Step 2.3). Decisions are related to development of the database to conduct counterfactual based micro (Step 3.1 and 3.2) and macro (Step 4.1 and 4.2) level evaluations. Activities include the set-up of data infrastructure for counterfactuals and development of procedures and protocols. Decisions relating to these activities have a strong impact on the effectiveness of the available data sources. Further, the maintenance of the database is important for ensuring the long term availability of data generated. Decisions in this evaluation phase are mainly related to increased work load or the kind of equipment (e.g. software) that is used in the analysis. The investment in development of a robust database and its maintenance could increase the effectiveness of the evaluation method and enable the use of the data base for future evaluations.

The application of the method (fourth phase of the evaluation cycle) uses the database developed to implement counterfactual based micro and macro-level analysis (Step 3.3 and 4.3). Analysis is based on the previous assessment. The suitability of the selected indicators based on the data availability is tested and adaptations are implemented (if required). Decisions are required about the mode of analysis and variations of the testing which directly influence the quality of the

evaluation results. Usually, this decision is related to an increased work load for the evaluator. The accuracy and quality of the analysis is directly influenced by the decisions in this evaluation step.

The final phase of the evaluation cycle refers to the Interpretation of results and conducting consistency checks (Step 3.4 and 4.4). The results of the analysis need to be communicated to the target group. Depending on the complexity of the analysis greater efforts could be required to 'translate' scientific results into understandable and unambiguous policy recommendations. Decisions are required regarding time spent for the evaluation, usually with associated investment in personnel and equipment, but innovation may offset those costs, such as in relation to communicating results.

Conducting consistency checks (Step 3.4 and 4.4) is essential to validate the results of the analysis and increase its robustness. Decisions have to be made on the mode of analysis for consistency checks. Costs arise due to increased staff time on consistency checking. Further, additional costs for equipment might be necessary, e.g. when the use of further statistical software is required. The quality of the results increases when sufficient time is spent on the communication and development of policy recommendations as well as the validation of the results through consistency checks. Thus, decisions in this evaluation step have a strong impact on the effectiveness of the evaluation approach.

In conclusion, in all evaluation phases decisions are required which will influence the cost-effectiveness of the evaluation approaches. This is particularly true of decisions at the outset of the evaluation cycle, thus in the first steps in application of the logic model, which have impacts on the overall effectiveness of the evaluation as they influence data generation, database development and applications of the evaluation method. However, good decisions at the outset (e.g. with respect to selection of indicators in Steps 1.1 and 1.2) cannot support good quality evaluation results if subsequent decisions in the evaluation process (e.g. with respect to the selection of counterfactual options in Step 2.3) inhibit the analysis. Thus, an appropriate level of resources can be expected to facilitate a successful evaluation.

5.2 Possible solutions for dealing with data gaps – impact on the cost-effectiveness of evaluation approaches

Data gaps constraint the effectiveness of direct environmental indicators and advanced methods. During the testing of the ENVIEVAL case studies, some partners experienced restrictions in the statistical analysis due to a lack of data access or data gaps. In those cases, the need for

improvements of the data environment is fundamental to facilitate the application of advanced methodological approaches. However, the impacts of data gaps on the effectiveness of indicators and methods need to be compared with additional cost of improved environmental monitoring programmes. This requires the consideration of different scenarios for future environmental monitoring programmes. Based on the results of the case study testing, three key types of scenarios could be derived:

- Additional efforts to increase the sample size and to improve the spatial coverage of the monitoring programme,
- Strategic sampling design of monitoring programmes exploring options to reduce monitoring efforts while at the same time to improve the spatial targeting of participants and non-participants
- Better integration of existing monitoring data from different sources or / and better integration of environmental monitoring data with farm structural data

Four case studies were selected to develop cost scenarios in order to show ways to optimize the resource use and facilitate the application of the tested evaluation approaches. Those scenarios show possibilities how to improve the cost-effectiveness of the tested evaluation approaches by changes in the data environment or access. The expected impacts on the related costs and on the performance of the evaluation approach are analysed.

The selected case studies cover different public goods (water quality, climate stability, biodiversity wildlife and landscape) in diverse country-specific conditions (Germany, Italy, Hungary and Scotland).

5.2.1 Water quality – the example of Lower Saxony

Scenario: Improved sample selection by developing a strategic sampling approach

*Through improved sample selection the coverage of all relevant measures with a sufficient sample size is ensured to enable robust counterfactual analysis using statistic-based evaluation methods***.**

The indicator ‘mineral N content in the soil in autumn’ (Nmin) resembles the nitrate that is potentially washed out into the groundwater during the winter period. It is addressing the need to reduce remaining mineral N in the soil after harvest and undesired mineralisation of N from the soil pool, and subsequent leaching into the groundwater. The Nmin indicator is addressed through agri-environmental measures and can be measured shortly after the measures implementation. Monitoring data is commissioned by the monitoring organisation (NLWKN) in water extraction areas in Lower Saxony. 20,000 soil samples are available for the years 2000 to 2006 and represent

sites with agri-environmental measures (AEMs) as well as sites without. For the assessment a pairwise comparison and regression analysis were used. Further information on the application of the indicator can be found in publications of the managing authority. Information is only available in German (NLWKN, 2010 and NLWKN, 2015).

Changes to data environment and data access

In the case study testing and in a previous study (Schmidt and Osterburg, 2010), it became clear that despite the large sample size not all sub-measure groups could be analysed due to small sample sizes for some measures. On the other hand, sample sizes of some sub-measure groups with known impacts were larger than necessary. This scenario aims to improve sample selection to enable a robust counterfactual analysis of all relevant measures. In this particular case, a more balanced coverage leads to reduced sampling size for some measures with very large sample sizes and secure effects (e.g. catch crops) and increased sampling size for other groups (e.g. restoration of extensive grassland). Therefore, through improved planning of sample selection and coverage of all relevant sub-measure groups the sample size can be decreased.

The implementation of a strategic sampling approach does not imply the use of additional data sources as Nmin values are collected for monitoring purposes in Lower Saxony. Generally, Nmin values could be used by the evaluator but the indicator is currently not utilized for the evaluation as the focus is on the CMEF indicator (GNB). During the case study testing it turned out the data of recent years is only available as an aggregated data set at the level of the water protection area. To apply this approach, micro-level data needs to be available to conduct an impact assessment. Thus, a revised arrangement of data access for RDP evaluation would enable access to micro-level data including all relevant sub-measure groups (14) with sufficient sample size for counterfactual analysis. Previously, only 10 sub-measure groups could be analysed with pairwise comparison. For an improved coverage of participants and non-participants for the different sub-measures using a strategic selection of relevant samples, a 1:3 matching algorithm should be used.

Impact on cost

This approach probably would be associated with a decrease of the cost considering the available sample size of 20,000 values for the year 2000 to 2006. The managing authority and/or the monitoring organization could benefit from the cost reduction. However, it is important to mention that autumn Nmin samples are collected for multiple purposes such as environmental monitoring and for educational reasons. Evaluation is only a secondary objective but when including evaluation needs in the consideration of sample selection, the cost-effectiveness of the resource use

would be enhanced. In this scenario we focus on the data requirements for the evaluation purposes. Thus, we consider the cost of the improved coverage in addition to the existing monitoring data set. Further, the integration of relevant measure combinations is considered as it further improves the performance of the evaluation approach. In the case of Lower-Saxony, no additional costs are related to the inclusion of measure combinations as the sample size of the four relevant combinations are sufficiently covered (Opt_1).

Further, we explore the cost for the case that the whole monitoring data has to be collected. This refers to the situation when this kind of monitoring data is newly established in another country. It shows the related cost for the transfer of this approach in another context. We consider two specific situations: first, the case when the main 14 sub-measure groups are considered (Opt_2) and second, the addition of relevant measure combinations (Opt_3).

The monitoring cost are based on the contracting cost for extensionists who collect the data (61.50 Euro per sample) as defined in the calculation basis of the Agricultural Chamber (LWK) of Lower Saxony (LWK Niedersachsen, 2014). The cost of 61.50 Euro include sample taking in 3 soil layers and the analysis of the sample in the laboratory.

Schmidt and Osterburg (2010) concluded that a minimum of 100 samples should be available for each AE sub-measure related to water quality objectives. For sub-measures with a secure effect, 50 samples might be sufficient. In this scenario we calculate the required sample size more generously with 110 samples for each sub-measure group as in the study of Schmidt and Osterburg (2010) 90% of the existing samples could be used for the analysis. 10% of the samples did not fulfil the criteria e.g. in consideration to the depths of the measurement (90 cm) or the timing of the sample taking (between October 1st and November 30th). Thus, in addition 10% more samples are considered in order to ensure the sample size of 100 is available for the analysis. Thus, a sample size of 1,540 measurements is considered for the participating sites.

As one site with a measure is compared to three similar sites without measures, a simplified estimation would consider 300 samples as a reference for each sub-measure group (4,200 samples). This would be the case if all sub-measures have different reference groups and no overlaps exist. In reality, this is rarely the case as different measures are implemented on similar sites and therefore the same reference can be used. Thus, the sample size can be decreased further if a strategic sampling approach considers the coverage of the main characteristics (e.g. main crop and soil conditions) of the participants to construct a suitable reference group.

In this example, most of the 14 sub-measures have overlapping reference groups as they are implemented on similar sites. Those reference groups are conventional land use, conventional arable land use and conventional grassland use. Only four measures have stand-alone reference groups but some of them overlap with the reference situation of the conventional (arable) land use (e.g. reduced row-width of corn is compared to conventional corn production). Only two of the measures have very specific reference situations with few overlaps. Greening of fallow land is compared with fallow land without greening and reduced pesticide application is mainly implemented for potato and barley production which is not largely covered in the conventional arable land-use situation.

To approximate the number of required samples for the reference group, a simple calculation is used. The distribution of the RD measures implementation is used to calculate the ratio of the main 20 crops for participating sites. For each crop, the measure with the highest appearance in the participant group is used and multiplied with 3 to detect the required amount of reference situations. This leads to the estimated sample size for non-participants of 2,115 samples. As in this analysis, also five soil quality classes referring to nitrogen leaching potential, the amount of organic fertiliser input (eight classes) as well as two years (2003 and other than 2003 as climatic conditions vary strongly) the combination of different characteristics increases the sample size in order to be able to cover as much specific situations as possible. The sample size could be reduced if the analysis would focus on the main and most frequent situations and thus, e.g. the most frequent classes of organic fertiliser or soil quality. Therefore, it is very difficult to estimate the exact number of required samples for non-participants. The reality is assumed to be between the two extremes of 2,115 samples or 4,200 samples. Thus, the average of the two values (3,158 samples) plus an addition of 10% is used for this scenario to simplify the assumption which leads to a sample size of 3,473 for non-participants.

Based on these assumptions, three options are identified for calculating changes in sampling size through strategic sampling.

Opt_1 refers to the cost of the sample in addition to the existing sample of 20,000 measures
Opt_2 refers to the total sampling cost for participants and non-participants for the 14 sub-measure groups
Opt_3 refers to the total sampling cost for participants and non-participants for the 14 sub-measure groups and four measure combinations.

The cost are considered for the whole period covered (7 years) and calculated for cost per year.

The case study testing used 20,000 samples that were collected within seven years, thus approximately 2,857 samples each year. Opt_1 presents the situation when the strategic sampling approach is added as a top-up to the existing set of monitoring data. Thus, only costs for the additional sampling of participants to reach the minimum requirement of 110 samples per sub-measure group are considered. 343 samples (49 per year) are necessary to achieve the required sampling size of participants in each sub-measure group. The inclusion of the four measure combinations does not lead to extra cost as the existing sampling size of each combination is higher than the minimum required sampling size.

With the identified 14 relevant sub-measure groups this leads to a sampling size of 5,013 samples for Opt_2. The strategic sampling approach would reduce the sampling size to 716 samples per year. When adding 4 important measure combinations (Opt_3), the sample size increases only for the participants as equal references can be used for the combinations and the single sub-measures. The sample size increases to 5,453 (776 per year).

Table 18 Overview monitoring cost in EURO

Activity / Cost types	Sample size	Baseline	Scenario: Strategic sampling	Additional cost
Sampling size for one year				
Opt_1 Existing data set with strategic sampling approach (top-up)	49	174,286	178,719	**3,014**
Opt_2 Strategic sampling for 14 sub-measures	716	0	44,045	**44,045**
Opt_3 Strategic sampling for 14 sub-measures + 4 measure combinations	779		47,911	**47,911**
Sampling size for whole period (7 years)				
Opt_1 Existing data set with strategic sampling approach (top-up)	343	1,230,000	1,251,095	**21,095**
Opt_2 Strategic sampling for 14 sub-measures	5,013	0	308,315	**308,315**
Opt_3 Strategic sampling for 14 sub-measures + 4 measure combinations	5,453		335,375	**335,375**

The consideration of the additional monitoring cost for improving the existing data set (Opt_1) is associated with a low increase of cost. The set-up of new monitoring systems cause higher cost but are lower than the cost of the existing monitoring system in Lower Saxony. This suggests that cost reductions of the existing monitoring programme are possible. However, it should be considered that this scenario only addresses the purpose of evaluation. In reality monitoring systems often fulfil multiple-purposes such as in Lower Saxony and evaluation is only one of them. The larger sample size might be necessary to meet other purposes. The integration of relevant measure combinations (Opt_3) is associated with a small increase of cost. As the programme period usually takes six years, annual data should be considered. Thus, the cost of the period of seven years are more relevant and a good approximation to the real cost.

Impacts on effectiveness

The improved sampling of the monitoring data for the Nmin indicator improves the performance level of the evaluation approach. The improvement of the effectiveness is assessed by using the criteria framework using seven judgement criteria which are related to the main evaluation challenges. The effected performance criteria are presented in the following table.

Table 19 Impacts on the performance of the evaluation method of a strategic sampling approach

Judgement criteria	**Previous performance level**	**New performance level & explanation**	
		Integration of 14 sub-measure groups (Opt_2)	**Integration of 14 sub-measure groups + 4 measure combinations (Opt_1 and Opt_3)**
Compatibility with local environmental and farm structuralcharacteristics	Medium	No impacts expected: performance level remains medium.	No impacts expected: performance level remains medium.
Timing of environmental impacts captured	Medium	No impacts expected: performance level remains medium.	No impacts expected: performance level remains medium.
Establishment of robust causal relationships	High	High Performance level remains high but improved coverage of more sub-measure groups improves the robustness of the causal linkages.	High Coverage is further improved as besides a better coverage of more sub-measure groups also measure combinations are considered. This improves the robustness of the causal linkages further.
Assessment of net impacts	Medium	Medium to High More robust impact assessment through further reductions in selection bias	High With the integration of measure combinations the net effect of single measures can be improved. Additional benefits of combined implementation of measures could be analysed.

Establishment of micro-macro linkages	Medium	High Improved robustness of results facilitates upscaling of micro level results to the macro level.	High Improved robustness of results facilitates upscaling of micro level results to the macro level.
Appropriateness of indicator(s) to capture complexity of environmental relationships	High	No impacts expected.	No impacts expected.
Unambiguous and understandable results and policy recommendations	High	No impacts expected.	No impacts expected.

A strategic sampling approach directed to fit the needs of the RDP evaluation improves the robustness of the causal linkages as more sub-measure groups could be covered by the analysis, and because sample sizes are standardized. The integration of measure combinations enhances the robustness of the results further. The strategic sampling approach reduces the selection bias which leads to a more robust net-impact assessment. With the integration of measure combinations the net effect assessment of single measures can be further improved and additional benefits and synergies of combined implementation of measures could be analysed. Moreover, a strategic sampling approach improves the representativeness of the data which facilitates the upscaling of the results to macro level. Thus, the performance level of the criteria 'establishment of micro-macro linkages' is enhanced to the high performance level. The other criteria are not directly affected by the strategic sampling approach.

5.2.2 Climate stability – the example of Emilia Romagna

Scenario: Measuring GHG emissions from agriculture with the carbon footprint at process level

Through improved utilisation of monitoring data and the benefits of additional monitoring data the application of elaborate statistic-based evaluation options is facilitated.

The Carbon Footprint approach has been implemented on different activities and sectors measuring the amount of carbon dioxide emissions that are directly and indirectly caused by an activity or accumulated over the life stages of a product at process level.

In the climate case study of the ENVIEVAL project, the analysis of the GHG emission at process level aims to contribute to measure the impact of an agricultural activity (e.g. productive process). The study is based on the analysis of the productive process and related RDP measures by the Emilia Romagna Region during the period 2007-2013. The case study wants to explore the impact of agri-environmental measures on the reduction of GHG emissions from agriculture.

Besides other secondary data sources, the assessment of carbon footprint at process level builds on the use of additional monitoring data that was collected by the evaluator. The cost scenarios build on the improved utilisation of this monitoring data collected at farm level and on the benefits of increasing the number of farm surveys at one or two points in time to facilitate the application of elaborate statistic-based evaluation options.

The scenario considers three options how the cost-effectiveness of this evaluation approach could be improved using the same type of farm level monitoring data:

- Opt_1) Re-processing monitoring data on cropping systems in order to use statistics-based evaluation options
- Opt_2) Additional ad-hoc survey on livestock systems in order to use statistics-based evaluation options
- Opt_3) Repetition of the estimation for period t+1 for implementation of Difference-in-Difference (DiD) options

Changes to data environment and data access

Existing data for the calculation of the carbon footprint is easily accessible for the evaluators as the managing authority has an interest to make data available for them. The first option (Opt_1) considers the re-processing of monitoring data on cropping systems in order to enhance the possibilities to use statistics-based evaluation options. It includes only the use of existing data sets and does not require any changes in data access. To use the improvements of the other two options (Opt_2 and Opt_3) additional data collection is necessary. Opt_2 refers to the conduction of additional surveys for livestock farms while Opt_3 aims for a repetition of surveys at time t1 for comparison with time t.

All scenarios are associated with a better definition of individual data. Thus, the coverage of participants and non-participants is improved.

Impact on cost

The additional cost for data collection has to be carried by the managing authority while the additional activities have to be covered by the evaluator. This implies a new contract between the managing authority and the evaluator. The increased working time for data processing and application of the evaluation method as well as the interpretation of results has to be covered by the evaluator.

As the use of Opt_1 does not require additional data collection, extra costs occur only for the additional activities of the evaluator. The increase in cost is associated with 20 additional working days for data processing for the use of existing data sets as well as the database development. Fifteen days are estimated to be necessary for the application of the counterfactual with statistics-based evaluation methods (such as PSM) and the analysis of the results.

An additional survey on livestock systems requires additional data collection. It is assumed that 200 visits to livestock farms are necessary and one survey costs 100 Euro. Thus, the survey would cost 20,000 Euro. For data processing 15 extra days are calculated and 5,000 Euro are spent for equipment for the database development. The conduction of the statistics-based evaluation method requires 10 additional working days.

The third option is associated with a repetition of the farm survey at another point in time in order to facilitate a difference-in-difference (DiD) analysis. To maximise the improvement of the evaluation approach (best case scenario), this is a top-up for Opt_2. This means that additionally to the improvements in Opt_2, the survey is repeated at a second point in time. Thus, the cost of Opt_2 can simply be doubled.

The impacts on the cost of the three scenarios are presented in the following table. It is compared to the cost of the baseline assessment which was conducted in the case study testing in WP6. The cells with coloured background indicate an increase in cost compared to the baseline.

Table 20 Overview of additional monitoring cost (compared to baseline) in EURO

Evaluation phases	Baseline assessment	Opt_1 Re-processing monitoring data	Opt_2 Additional ad-hoc survey on livestock systems	Opt_3 Repetition of surveys at another point in time
Evaluation design	14,420	14,420	14,420	28,840
Data generation	83,080	83,080	107,080	214,160
Database development	21,780	21,780	38,280	76,560
Application of method	9,410	13,260	15,780	31,560
Interpretation	5,830	8,280	9,960	19,920
Total cost	**134,520**	**146,820**	**185,520**	**371,040**
Additional cost		**12,300**	**51,000**	**236,520**

The comparison shows that with an increased labour demand and the need for additional data collection in Opt_2 and Opt_3 the cost increase. The estimated cost of Opt_1 is 12,300 Euro higher than the baseline as more labour is needed to apply the statistics based evaluation method as well as to interpret the results. The costs are 9 % higher than in the analysis of the case study testing.

Opt_2 is associated with additional data collection which has the biggest impact on the increase of the cost. The improved coverage of livestock farms in the survey is estimated to have 51,000 Euro higher cost than the baseline assessment. This refers to an increase of the cost of 35 %.

The repetition of the survey is of course associated with high additional cost. Compared to the baseline scenario, this would lead to an increase of the total cost by 236,520 Euro.

Impacts on effectiveness

The improvement of existing monitoring data enhances the performance level of the evaluation approach. The improvement of the effectiveness is assessed by using the criteria framework having seven judgement criteria which are related to the main evaluation challenges. The effected performance criteria are presented in the following table which compares the performance of the three options of the scenario.

Table 21 Impacts on the performance of the evaluation method (Carbon footprint, Italy)

	Previous performance level	**New performance level & explanation**		
	Baseline assessment	**Opt_1 Re-processing monitoring data**	**Opt_2 Additional ad-hoc survey on livestock systems**	**Opt_3 Repetition of surveys at another point in time**
Compatibility with local environmental and farm structural characteristics	Medium	No impacts expected.	Medium/high Increase sample size for livestock systems with a better representativeness of the variety of regional productive systems	Medium/high Increase sample size for livestock systems with a better representativeness of the variety of regional productive systems
Timing of environmental impacts captured	Low	No impacts expected.	No impacts expected.	High Integration of a second point in time (t1)
Establishment of robust causal relationships	Medium	Not relevant for the selected options	Not relevant for the selected options	Not relevant for the selected options
Assessment of net impacts	Medium	Medium/high Increases the chance to use statistics-based evaluation options for the analysis of cropping systems.	Medium/high Increases the chance to use statistics-based evaluation options for the analysis of livestock systems.	High Increases the chance to use statistics-based evaluation options based on before-and-after comparison, taking into account possibly DiD approaches.
Establishment of micro- macro linkages	Medium	No impacts expected.	Medium/high Allows for a better representativeness of livestock systems	Medium/high Allows for a better representativeness of livestock systems
Appropriateness of indicator(s) to capture complexity	High	Not relevant for the selected options	Not relevant for the selected options	Not relevant for the selected options

of environmental relationships				
Unambiguous and understandable results and policy recommendations	High	Not relevant for the selected options	Not relevant for the selected options	Not relevant for the selected options

The re-processing of monitoring data on cropping systems to be able to apply statistic based evaluation options (Opt_1) is expected to improve the net impact assessment but only for the cropping system analysis. The other two options also have a positive influence on this criterion.

The second option which integrates data on livestock systems through an ad-hoc survey (Opt_2) is expected to have an impact on three performance levels. With the inclusion of livestock farms, the increased sample coverage improves the level of representativeness and the compatibility with local environmental and farm structural data. Further, the integration of livestock systems allows a better representativeness and possibly improves the micro-macro linkages of the approach.

The more advanced scenario (Opt_3) which integrates another time period (t+1) in the assessment has an impact on four performance criteria. Through the integration of another time period, the timing of environmental impacts is improved to a high performance level. The approach also achieves a high performance level in terms of assessment of net impacts, due to the chance to compare groups before-and-after and to use elaborate-statistics evaluation methods such as Propensity Score Matching with Difference in Difference approach.

5.2.3 Biodiversity (FBI) – the example of Hungary

Scenario: Improved access to existing biodiversity data for counterfactual analyses

The aim of the scenario is to enable improved access to existing spatial explicit data to enable a more targeted sampling and thereby improve the counterfactual analysis.

The CMEF defines the Farmland Bird Index (FBI) impact indicator addressing to evaluate the impacts of the RD measures on the changes in biodiversity. FBI is a composite index that measures the rate of change in the abundance of common bird species at selected sites, i.e. relative abundance. The indicator summarises species' trends in 25 European countries. Each country selects the common species most representative for the respective habitat from a list of 37 indicator species that are common and characteristic of European farmland landscapes. Birds are a good indicator for wildlife biodiversity as they highly depend on biodiversity elements in habitats and are responsive and sensitive to environmental change (EC, 2012). Population trends are derived from the counts of individual bird species at census sites and modelled as such over time.

Changes to data environment and data access

The scenario for the FBI indicator is related to the improved access of existing data for the counterfactual analysis. Currently, the spatial distribution of the contracted parcels under different RD measures (eg. agri-environmental schemes) is usually only provided to the evaluators by the end of the programming period for conducting ex-post evaluations. By including information on the spatial distribution of AEM contracted areas in the beginning of the contracting period, the data collection could be adapted to fit better the evaluation needs. Currently, the so called 'contracted' survey squares are covering only a relatively low share of the contracted parcels. When the spatial distribution of the contracted parcels is known, the survey spots can be placed at the exact geographical places throughout the country directly targeting participants and non-participants. The approach would improve the counterfactual coverage as the data would be available for the whole programme period.

Additionally to this improvement, the use of the baseline data of the FBI for micro-level analysis is explored (Opt_2). The 'number of farmland birds' is based on the raw data of the FBI calculations. The baseline data of the FBI was analysed in a temporal scale. In each survey square (2.5 km^2), 15 survey points exist. The number of birds is counted within a 100m radius from the centre of the spot. Each survey point covers approximately 3 hectares of observed territory that are used for this analysis. Each point represents the biodiversity at parcel level which compared with the additional attributes and uptake characteristics, provides a good possibility for impact assessment. No new data is necessary as the baseline data of the FBI is used for the analysis. The raw data of the common bird monitoring programme is needed to be able to carry out micro level assessments. Besides using LPIS data for identifying participant and non-participant survey groups, survey spots were further grouped by the rate of natural areas recognized within the spots based on land cover data. The quality of land cover data sources may influence the final results. Therefore data is needed in a relatively good resolution.

Impact on cost

The improvement of the performance of the evaluation approach is not associated with an increase of cost as the number of sampling squares does not change. 200 survey squares are included in the assessment, which cover the whole country representatively. Approximately five hours are necessary for the survey of one square. As surveys shall be conducted twice a year, 10 hours are allocated for the monitoring of one survey square. Data processing for each square takes approximately five hours. The salary of the monitoring staff is estimated to be 100 Euro per day. Please note that most of the field work for monitoring the FBI in Hungary (approximately 90 %) is

carried out by volunteers. Thus the calculation of the monitoring cost is based on assumptions if volunteers are not available and is not reflecting the real world situation. Travel costs are based on the assumptions that 30 Euro are spent on the round trip for one survey square per year. Indirect costs are calculated with 20 % of the direct cost. Estimated costs may differ among member states.
To add micro-level analysis of the raw data set, only the use of existing data sets are required. Costs occur for the additional work load that is required to conduct the micro level analysis. For data processing additional five working days are expected to be necessary for monitoring staff with a salary rate of 100 Euro per day. For the comparison group design GIS expertise is needed. It is expected to require five extra working days of scientific staff (300 Euro per day). Also the application of the method is expected to require five extra working days of a researcher.

Table 22 Overview monitoring cost of the farmland bird index (FBI) in Hungary

Activities	Baseline scenario (Opt_1)	Micro level analysis (Opt_2)
Field work	37,200	37,200
Data processing and storage	15,000	17,000
Application of method	3,600	5,100
Total cost	**55,800**	**59,300**
Additional cost		**3,500**

The estimation of the cost for the monitoring activities and data processing show that data collection is the main cost source for the use of the FBI indicator. For Opt_1, no additional costs are necessary for the improvement. The integration of micro level analysis (Opt_2) in the assessment is associated with additional work load which leads to an increase of cost about 3,500 Euro.

The costs in the table provide information on the monitoring of the FBI for one year. As a programme period takes six years, the estimated monitoring cost for one evaluation period are approximately 334,800 Euro and 355,800 Euro respectively.

Impacts on effectiveness

The improved availability of the biodiversity data for counterfactual analysis improves the performance level of the evaluation approaches. The improvement of the effectiveness is assessed by using the criteria framework having seven judgement criteria which are related to the main evaluation challenges. The effected performance criteria are presented in the following table

Table 23 Impacts on the performance of the evaluation method (Farmland Bird Index – Hungary)

Judgement criteria	Previous performance level	New performance level & explanation	
		Opt_1 Improved spatial coverage	Opt_2 Improved spatial coverage + micro level analysis
Compatibility with local environmental and farm structuralcharacteristics	Medium	No direct impacts expected	Medium Micro level approach compatible with parcel level.
Timing of environmental impacts captured	High	No direct impacts expected	No direct impacts expected
Establishment of robust causal relationships	High	High With setting up better with/ without comparisons the establishment of causal relationships may be further improved.	High Assessment at micro level may provide even better insights of the impacts of the measures than at macro level.
Assessment of net impacts	Medium	High The improved coverage of participants makes the assessment of net impacts more robust. A more balanced and targeted sample selection improves the counterfactual analysis and increases the significance levels.	High Additional insights at micro level may improve the net impact assessment further as impacts of AE measures on local biodiversity may be better explored.
Establishment of micro-macro linkages	Medium	No direct impacts expected	High New method may help to cross-check the results of macro level assessment and improves the micro-macro linkage.
Appropriateness of indicator(s) to capture complexity of environmental relationships	High	No direct impacts expected	No direct impacts expected
Unambiguous and understandable results and policy recommendations	High	High With a more precise assessment of impacts a better understanding of RD measures is feasible which could further improve the quality of policy recommendations.	High With a more precise assessment of impacts a better understanding of RD measures is feasible which could further improve the quality of policy recommendations.

The more targeted monitoring is expected to have impacts on three performance criteria. Through the introduction of more robust counterfactuals, the establishment of causal relationships could be further improved although the approach already received a high performance level. Further, the net impact assessment could be improved as a better coverage of participants allows are more robust analysis which could lead to a higher performance level (medium to high). By considering participation in AE programmes, the sample selection is more targeted and the ratio of survey spots for participants and non-participant is more balanced. This will lead to higher significance levels and a more robust net-impact assessment. In addition to the improved monitoring of Opt_1, the integration of micro level analysis improves the performance of five judgement criteria.

Additional effects are the improvement of the coverage of environmental characteristics. Further, through the integration of the micro level analysis, the establishment of causal linkages and the net-impact assessment is further improved. The combination of micro and macro-level analysis improves the micro-macro linkages. Also, the quality of policy recommendation could by improved by using a more precise assessment of the environmental impacts in both options.

As the improvement of the performance level of the tested evaluation method for the FBI indicator in the Hungarian biodiversity wildlife case study is not associated with any additional cost, it is recommended to use this possibility to improve the impact assessment of this indicator. This would improve the cost-effectiveness of the evaluation approach by achieving a higher performance level without additional cost. Better cooperation between implementing agencies and monitoring bodies is recommended from the beginning of the programming period based on the evaluation plans obligatory for member states during the planning of Rural Development Programmes. The integration of the micro-level analysis of the number of farmland birds is associated with a relatively low increase in cost while the effectiveness of the approach is further improved.

5.2.4 Landscape – the example of Scotland

Scenario: Improved availability of landscape (monitoring) data

This scenario analyses the integration of remote sensing (RS) data, such as SENTINEL 2, into RDP evaluations to improve the data base on land cover change.

A considerable body of scientific literature exists on the assessment of changes in landscapes (i.e. interventions) expressed in terms of changes in metrics. The landscape metrics approach to the provision of an impact indicator considers the use of data on land use and changes in land use due to the uptake of RDP Measures. There is a range of indicators that are commonly used to analyse changes in landscape structure. These generally draw on metrics from the fields of landscape ecology and spatial analysis, e.g. topological, network, dispersion, shape, and indices to express metrics by functional, administrative or other forms of spatial units (e.g. regular grids). In this approach a selection of such indicators are explored for their suitability as RDP impact indicators.

Changes to data environment and data access

This scenario of the Scottish landscape case study considers the integration of remote sensing (RS) data, such as SENTINEL 2, into RDP evaluations to improve the data base on land cover change. Currently, CORINE land cover data are updated every 6 years for monitoring land cover change

however the monitoring cycle does not align with the RDP evaluation cycle, which means that although they are suitable for the evaluation of change at sub-parcel level due to their temporal misalignment they are not suitable for a reliable impact assessment. In the landscape case study, IACS land data are used to align the data with the RDP. However these data are only available at parcel level, while the impact of RDP is commonly delivered through sub-parcel changes. SENTINEL 2 has recently launched and is going to be a source of RS data, which will be used for monitoring land cover with a high temporal and medium spatial resolution. The Sentinel 2 land monitoring data is covering the Earth's land surface globally and is updated every 10 days using one satellite or every 5 days using two satellites (ESA, 2015). This will lead to improved land cover/land use data. The data will cover continuous spatial areas and include both participants and non-participants, which through an overlay can be integrated.

These data are new and are currently in the process of coming available for use. In the coming years the intention is that land cover data derived from these data will become more widely available. This data would be highly suitable for RDP evaluations.

Impact on cost

The cost of making this data available lies with EU (through Copernicus project). Depending on the suitability of the data made available in coming years, there may be additional cost for the evaluators/monitoring authorities. However, the land-cover product is freely available.

The main uncertainty is the format in which the data is going to be available. If the data is provided as a land cover product that is ready to use for the evaluation of RD programmes, the integration of RS data does not generate increased cost of data use or increased workload. The data would improve the CORINE database and provide benefits for the assessment of land cover change. However, if the remote sensing data is provided as raw image data, additional data processing will be required to classify the data before it can be used for RDP evaluation. These costs are for data processing as well as for additional high end computing processing power and storage facilities. These extra costs can impede the use of SENTINEL 2 data for RDP evaluation. Due to the uncertainties of the available data format both options are included in the assessment.

An evaluation using raw Sentinel 2 data may require the use of a single or multiple scenes. In case of an assessment that can be conducted using a single scene (regional level), the additional costs are related to the selection and classification of the scene and some additional computer processing power. The use of multiple scenes for the evaluation will demand high end computer processing and storage facilities as well as additional procession costs to integrate multiple scenes, which

requires 'stitching' of scenes. Also more time is required for the processing of the data. As the data is not available yet and thus no experience with the integration of the data in the RDP evaluation exists the following table is based on very rough estimations. The example can only provide an indication that the costs will increase.

The inclusion of the land cover product is not associated with any increase of the cost. It is equal to the cost of the evaluation approach tested in the ENVIEVAL case studies. When SENTINEL 2 data is integrated as raw data but only one scene (region) is included, an extra four days is expected for the assessment and downloading of the scene and five extra days for the processing of the satellite data and validation exercises. The increased demand for high-end processing and storage facilities is expected to increase to 3,000 Euros. For the integration of several scenes to cover a programme or national level is estimated to require 4 extra working days of the assessment and downloading of the scenes and 10 additional days for the processing and validation of the data at national level (multiple scenes). As extra high end processing and storage facilities are needed, the cost for computing power would increase about 20,000 Euro.

Table 24 Overview of additional cost of SENTINEL 2 data as product compared to raw data

Activity / Cost types	Baseline/ Land cover product	Raw data (regional)	Raw data (national)
Labour cost	9,478	12,313	16,723
Equipment	2,100	3,000	22,100
Travel cost	700	700	700
Indirect cost	9,100	12,250	17,150
Total cost	**21,378**	**28,263**	**56,673**
Additional cost		**6,885**	**35,295**

The comparison of the three possibilities of data format and availability shows that the first option, to receive products that fit the evaluation needs, is preferential as it is associated with the lowest cost. The cost of the raw data depends strongly on the amount of scenes that needs to be used for the assessment. If only one scene is used to cover the regional level, the cost would increase about roughly 7,000 Euro compared to the first option. The integration and combination of several scenes increases the work load and particularly the demand for improved computing power has a strong effect on the cost.

Impacts on effectiveness

The integration of high resolution remote sensing data (SENTINEL 2) improves the performance level of the evaluation approach. As the data set contains the same information for all three options of the data format and availability, the impacts on the performance are equal. The effected performance criteria are presented in the following table.

Table 25 Impacts on the performance of the evaluation method of the integration of SENTINEL 2

	Previous performance level	**New performance level & explanation**
Compatibility with local environmental and farm structural characteristics	High	High The data based on 10m raster is able to provide land cover data at sub parcel level. For the method this is important because it means that patches of similar land use /land cover can be identified more accurately hence the landscape structures .
Timing of environmental impacts captured	Medium	High SENTINEL 2 revisits the same place every 5 days. While weather conditions (cloud cover) may make the images useless, the frequency is high enough to generate a usable mosaic image of areas with regular cloud cover.
Establishment of robust causal relationships	Medium	High The increased detail in land cover data (sub-parcel) means that it is more likely that implementation of a measure can be 'observed' through the RS data which is a good basis for robust causal relationships between RDP measure and impact on landscape.
Assessment of net impacts	Medium	No impacts expected
Establishment of micro- macro linkages	High	High This does not change with the improvement of the data because this is inherent to the method, however the results will have been enhanced due to the higher accuracy of the land cover data.
Appropriateness of indicator(s) to capture complexity of environmental relationships	Medium – High	No impacts expected
Unambiguous and understandable results and policy recommendations	Medium - Low	No impacts expected
Compatibility with local environmental and farm structural characteristics	High	High The data based on 10m raster is able to provide land cover data at sub parcel level. For the method this is important because it means that patches of similar land use /land cover can be identified more accurately hence the landscape structures .
Timing of environmental impacts captured	Medium	High SENTINEL 2 revisits the same place every 5 days. While weather conditions (cloud cover) may make the images useless, the frequency is high enough to generate a usable mosaic image of areas with regular cloud cover.
Establishment of robust causal relationships	Medium	High The increased detail in land cover data (sub-parcel) means that it is more likely that implementation of a measure can be 'observed' through the RS data which is a good basis for robust causal relationships between RDP measure and impact on landscape.
Assessment of net impacts	Medium	No impacts expected
Establishment of micro- macro linkages	High	High This does not change with the improvement of the data because this is inherent to the method, however the

		results will have been enhanced due to the higher accuracy of the land cover data.
Appropriateness of indicator(s) to capture complexity of environmental relationships	Medium – High	No impacts expected
Unambiguous and understandable results and policy recommendations	Medium - Low	No impacts expected

The SENTINEL data provides high resolution data at sub parcel level and facilitates the comparison of patches with similar characteristics. This would further improve the compatibility with local environmental characteristics. The high frequency of data collection (every five days) enhances the timing of environmental impacts captured compared to an annually updated data set. The performance level increases from medium to high. The more detailed land cover data also improves the establishment of robust causal relationships and enhances the performance to a high level. Although the micro-macro linkage is also covered well with the current approach, due to the higher accuracy of the land cover data it would be further improved.

5.2.5 Synthesis of the tested cost scenarios

The four scenarios show how the cost-effectiveness of evaluation approaches can be improved. All scenarios aim to increase the chance to use statistics- based evaluation options. The cost scenarios of the Hungarian, German and Italian case studies deal with the improvement of monitoring data while the Scottish case considers the integration of improved remote sensing data. As the latter is not related to the improvement of monitoring data, it is not included in the comparative analysis of the scenarios. The synthesis of the scenarios results focuses on the three approaches related to the improvement of monitoring data.

As the conditions of data access and data environment vary between countries and public good, the examples show very different effects on cost and effectiveness of the tested evaluation approaches. Table 26 compares the effects on the cost and the performance of the evaluation approaches of the developed scenarios. The comparison includes the water quality case study in Germany (WQ-DE), the Italian climate stability case study (CL-IT) and the biodiversity wildlife case study in Hungary (BW-HU). For the German cost scenario, only the first option is included as this is related to the improvement of the existing data set which was available for the case study testing.

The scenarios address the improvement of monitoring data through different adaptations in the data environment. The following figure shows the interlinkages of the cost scenarios' impacts with the evaluation phases.

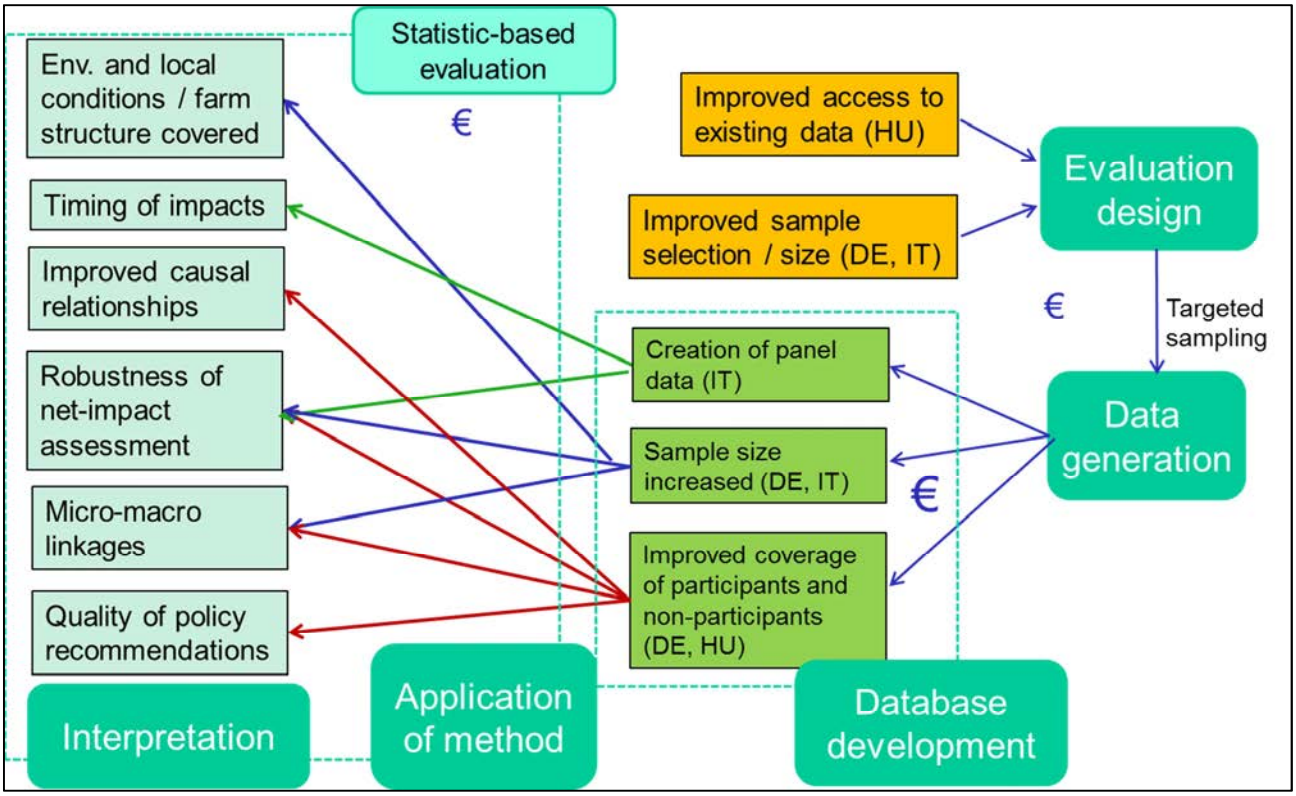

Figure 14 Overview of scenario impacts in the evaluation cycle

The scenarios influence the first phase of the evaluation process, namely the evaluation design, through improved access to existing data sets or the improved sampling design. Those enable a more targeted sampling approach that covers sufficient values for participants and non-participants to apply statistic-based evaluation methods. The improved planning has an impact on the data generation step. It can be associated with additional data collection as sample size is increased e.g. by covering multiple comparison groups or creating panel data, or rearrangement of samples to improve the coverage of participants and non-participants without increasing the total amount of samples. This improved database targets the application of statistic-based evaluation approaches to improve the effectiveness of the evaluation approach. In the case of the three selected cost scenarios, six effectiveness criteria are influenced by the improvements. Panel data improves the timing of environmental impacts by introducing a second point in time. The increased sample size with a better coverage of the comparison groups targets an improved coverage of environmental and local conditions and farm structures and improves the micro-macro linkages as the representativeness of data is improved. The improved coverage of participants and non-participants is expected to improve the establishment of robust causal relationships and the micro-macro linkage. Further, the quality of policy recommendations can be improved. All three

adaptations of the data base improve the robustness of the net-impact assessment which is essential for the conduction of impact assessments.

Naturally, the improvements cause cost. For the adaptation of the evaluation design and the facilitation of the application of statistic-based evaluation methods, the increase of cost is associated with additional workload of the staff. The improvements in the data generation step are often associated with additional data collection which usually causes higher cost. The cost and the impacts of the cost scenarios have to be compared to identify the impact on the cost-effectiveness of the evaluation approaches. Table 17 presents the comparison of the effects on the cost and performance of the evaluation approaches.

The cost of the case study testing present the baseline cost. The changes of cost associated with the implementation of the scenario are expressed in absolute and relative terms. The effectiveness is measured by counting the number of performance levels that are influenced by the application of the scenario. The number of improved performance level is also presented as the degree of impact can differ (e.g. change from low to medium level or low to high level). If a performance level is improved without reaching a higher performance level, the improvement is considered to be half of a performance level (0.5). The cost-effectiveness ratio is expressed as cost for the improvement per performance level.

Table 26 Comparison of the impacts on cost and effectiveness of the scenarios

Scenario	Hungary - FBI		Italy – Carbon footprint			Germany - Nmin
	Opt_1	Opt_2	Opt_1	Opt_2	Opt_3	Opt_1
Baseline cost	55,800		134,520			1,230,000
Additional cost (absolute)	-	3,500	12,300	51,000	236,520	21,095
Additional cost (share)	0%	6%	9%	38%	176%	2%
Number of improved effectiveness criteria	3	5	1	3	4	3
Number of improved performance levels	2	3.5	0.5	1.5	4	2.5
Increase of cost per performance level (%)	0%	2%	18%	25%	44%	0.7%

As the case study scenarios refer to different public good, countries and evaluation approaches, the impacts on the cost and effectiveness cannot be directly compared, and show high variations. The comparison is restricted to different options within each case study.

What can be seen in the Hungarian and the Italian cases is that with increasing cost the performance of the evaluation approaches increases. Opt_2 and Opt_3 of the Italian scenario imply additional data collection which has a strong impact on the increase of cost. The German case is also associated with additional data collection. However, due to the good data availability in the baseline scenario, the increase of cost is very low.

Scenarios that are not associated with additional data collection experience a lower increase of cost. In Opt_1 of the Hungarian case study, an increase in the effectiveness of the evaluation approach is not associated with any increase in cost as only the integration of an existing additional data set is recommended. The cost of the evaluation approach would remain the same. Also in Opt_2 of the Hungarian case and Opt_1 of the Italian case the improvement is not associated with additional data collection, thus the cost increase is relatively low. Those scenarios show that improvements of the performance of evaluation approaches can be often achieved with little efforts. The collection of additional data is often associated with increased cost but increase the effectiveness of evaluation approaches further.

To validate the importance of the improvements of the evaluation approaches, the impacts of the scenarios on the performance of the evaluation methodologies are compared to the stakeholder priorities that were identified in national stakeholder workshops (see section 3.2.2 Participatory assessment and validation and section 4.2.1 Defining weights for the criteria of the performance assessment). The participants in the workshops validated and assessed the importance of the effectiveness criteria of the developed framework for the performance assessment. Table 27 shows the average of the identified stakeholder priorities in the partner countries that developed cost scenarios. The two judgement criteria with the highest priority in each partner country are indicated in green.

Table 27 Stakeholder priorities of the national stakeholder workshops

Judgement criteria	Hungary	Italy	Germany
Compatibility with local environmental and farm structural characteristics	0.7	2.4	2.2
Timing of environmental impacts captured	3.8	1.9	1.7
Establishment of robust causal relationships	3.1	2.3	2.8
Assessment of net-impacts	1.6	1.8	3.0
Establishment of consistent micro-macro linkages	1.3	1.9	1.7
Appropriateness of indicator(s) to capture complexity of environmental relationships	2.4	2.6	1.5
Unambiguous and understandable results	2.1	2.3	2.2

The table shows that stakeholder priorities vary between different countries. Thus, the comparison has to be conducted for each partner country-specific set-up separately. The improved performance assessment of the cost scenarios is now compared to the priorities of the evaluators and managing authorities in each country. The table below shows the performance of the cost scenarios, indicating the level of performance with low, medium or high. The criteria that are highlighted with green colour are those criteria that are improved by the application of the cost scenario compared to the baseline situation in the case study testing. The two most important performance levels that were identified in each national stakeholder workshop for the respective case study are indicated in red colour.

Table 28 Comparison of results of the cost scenarios with stakeholder priorities of national workshops

Judgement criteria	Hungary - FBI		Italy – Carbon footprint			Germany - Nmin
Scenario	Opt_1	Opt_2	Opt_1	Opt_2	Opt_3	Opt_1
Compatibility with local env. and farm structural characteristics	Medium	High	**Medium**	**Medium/ high**	**Medium /high**	Medium
Timing of env. impacts captured	**High**	High	Low	Low	High	Medium
Establishment of robust causal relationships	**High**	**High**	Medium	Medium	Medium	**High**
Assessment of net impacts	High	High	Medium/ high	Medium/ high	Medium /high	**High**
Establishment of micro-macro linkages	Medium	Medium	Medium	Medium/ high	Medium /high	High
Appropriateness of indicator	High	High	**High**	**High**	**High**	High
Unambiguous and understandable results and policy recommendations	High	High	High	High	High	High

The comparison of the impacted judgement criteria and the stakeholder priorities show overlaps in each cost scenario. For the Hungarian case, the establishment of robust causal relationships is improved with the adaptations of the cost scenario. This criterion was given a high priority of the Hungarian stakeholders at the national workshop. The improvement of the carbon footprint approach in Italy meets the stakeholder's priorities by enhancing the compatibility with local environmental and farm structural characteristics in Opt_2 and Opt_3. The German case study scenario meets both criteria that received a high priority of the national stakeholder. Thus, it can be concluded that the improvements of the cost scenarios are highly relevant for the stakeholders in each country-specific situation.

5.3 Recommendations for the selection of evaluation approaches by the end-user under consideration of relative costs

The monitoring cost scenarios of selected ENVIEVAL case studies present possibilities how the cost-effectiveness of environmental evaluations can be improved by changing conditions of data access or data availability. The examples show that improvements can be achieved with a low increase of cost. Small efforts such as the integration of alternative existing data sets or a more

detailed analysis of available data can improve the effectiveness of evaluations. The costs increase usually further when additional data collection is necessary as this is time consuming and costly. However, with additional data collection the data base can be improved which effects the performance of the evaluation approaches.

It is recommended to include the data requirements and evaluation needs from the beginning of the development of a RD measure. The monitoring system should be jointly developed to be able provide the required data for conducting a sound environmental evaluation. Also, result-based schemes could be beneficial as the monitoring and control of the measure is a part of the measure's implementation.

The developed cost scenarios reflect specific conditions of data availability and access in the particular case study region. Therefore, the transferability of the improved availability of monitoring data to other countries needs to be explored. For the Hungarian biodiversity wildlife case study, it would be interesting to assess if this scenario would be suitable for application in the Lithuanian case study area as the farmland bird index is wildly used across member states. The transferability of the estimated cost and efforts of the improved evaluation through better data access need further validation.

The Scottish cost scenario represents an exceptional case within our cost scenarios. While the other cases deal with the improved access and availability of exiting monitoring schemes, the improvement of the Scottish evaluation approach is associated with improved remote sensing data that will be available in the future. As this data is not yet available, the scenario includes three options as the data format in which data will be provided is uncertain. This allows the coverage of different possibilities of data availability. As the impacts on the performance criteria are the same, the cost increase is not associated to an increase in effectiveness but is influenced by the provided data format and the coverage of the evaluation (regional or national level). It can be concluded that the integration of SENTINEL 2 data is favourable, as the effectiveness is improved. To what an extent it is associated with an increase of cost can only be seen when the data format of this new data set is known.

This shows that foreseen development of existing and new databases would be beneficial. Further, monitoring systems usually have multi-purposes such as monitoring and evaluation or educational purposes. It is recommended to consider these multiple purposes when a monitoring system is established to fulfil different data requirements. This would enhance the cost-effectiveness of evaluation approaches and facilitate the application of statistic-based evaluation options. In some

cases, such as the German cost scenario (Nmin indicator), the sampling size could be reduced when a targeted approach is used. This could lead to cost reductions and improve the cost-effectiveness of the monitoring programme. Naturally, the multiple purposes of monitoring data still need to be considered.

6 Conclusions

Within the ENVIEVAL project a structured approach is tested to assess the cost and performance of the evaluation approaches for different public goods in the case studies. The identified costs of the required resources were collected for each tested evaluation approach. The cost templates could help evaluators to plan and control evaluation cost in a structured way and to identify the main drivers of cost. The comparison of costs of evaluation approaches remains challenging although the detailed assessment of cost helps to show the drivers of cost for each evaluation approach. Comparability is further limited due to different conditions in the partner countries (e.g. different data access and expertise for statistical analysis) and evaluation agencies. This shows that the mere comparison of cost of evaluation approaches is not sufficient. But what is important is to raise the awareness of what suitable and advanced evaluation including adequate environmental programmes cost. This has also been particularly highlighted in the stakeholder workshops. It is also important to consider the effectiveness of the approaches in order to get a holistic valuation of the cost-effectiveness of evaluation approaches.

The summary of the performance assessment of the tested evaluation approaches highlighted how the different stakeholder priorities affect the interpretation of the results and ultimately the selection of the approach for environmental impact evaluations of RDPs. The results of the effectiveness or performance assessment can be differently interpreted depending on the set of priorities attached to the judgement criteria and the final decision which evaluation approach to select often depends on the particular priorities of the stakeholders. The final selection revolves around an inspection of the performance assessment considering the relative costs of the different approaches as well as specific circumstances, preferences and abilities of the end-user (stakeholder). It is however important that a consistent framework with clearly defined criteria and performance or impact levels is used.

The identification of stakeholder priorities and their different weights for judgement or effectiveness criteria of evaluation approaches is important for an ex-ante assessment of the potential contributions of possible approaches, informing the selection of evaluation approaches.

The explicit consideration of different stakeholder priorities also contributes to a better understanding to what extent the applied evaluation approaches have delivered the required results, addressed existing evaluation challenges and helps to identify the need for further improvements in both the data infrastructure and methodological development. The development of the conceptual framework with a set of quality and judgement criteria as well as performance levels provides the basis for a robust and sound assessment of the effectiveness of evaluation approaches. The framework developed in the ENVIEVAL project has attempted to fill the gap of a lacking framework and provides a starting point for further improvements of effectiveness assessments of environmental evaluations of RDPs.

Detailed assessments of the performance of the tested evaluation approaches using the framework developed in section 3.2 have been reported in the case study summary reports in Deliverable D6.3. Here only a short summary of the performance matrix of the tested evaluation approaches was provided. The high performance levels for the 'Establishment of causal relationships' and the 'Appropriateness of indicators and methods to capture the complexity of environmental relationships' highlight the emphasis of the public good case studies on contributions to address indicator gaps and contributions of advanced modelling approaches for dealing with the complexity of public goods (see also the discussion section of Deliverable D6.3). At the opposite end, only 3, respectively 4, tested evaluation approaches achieved a high performance level for the criteria 'Establishment of consistent micro-macro linkages' and 'Assessment of net-impacts', which reflects the severity of the methodological challenges underlying those two criteria as well as the large data requirements of evaluation approaches able to address these challenges.

During the evaluation process different decisions along the steps of the logic model influence the cost and effectiveness of the evaluation approaches. It can be concluded that in all evaluation steps decisions have to be made that influence the cost-effectiveness of the evaluation approaches. Particularly decisions in the beginning of the evaluation process and related to data availability have impacts on the overall effectiveness of the evaluation as they influence data generation, database development and the application of the evaluation method. However, good decisions in the beginning cannot provide good evaluation results if later decisions in the evaluation process inhibit the analysis. Thus, a balanced and considerable resource use could help to facilitate a successful evaluation. This is hampered by the limited funding and time restrictions that are available for evaluation. A realistic cost calculation and a targeted evaluation could help to overcome these restrictions.

The implementation of the monitoring cost scenarios for selected case studies of the ENVIEVAL project show that an improvement of the effectiveness of evaluation approaches can be achieved with relatively low cost, at least if one puts the additional cost into the context of the overall RDP budget. Also, small efforts such as the integration of alternative existing data sets or a more detailed analysis and processing of available data can already improve the effectiveness of evaluations. Further cost savings can be achieved by embedding additional data collection, or more generally, environmental monitoring for the evaluations of RDPs into a multi-purpose monitoring system.

If additional data collection is necessary to improve the evaluation method, cost are usually higher as data collection is costly and requires more efforts. The improvements either enable the use of advanced counterfactual methods or increase the cost-effectiveness of using those methods. Advanced counterfactual methods are crucial to be able to assess net impacts of RD measures. Improved monitoring data is also needed for the assessment of synergies between measures to enable the analysis of multiple comparison groups. Further, the improvements meet largely the stakeholder priorities identified in national stakeholder workshop in the partner countries. This is a validation that the cost scenarios address the main evaluation challenges of the particular case study setting.

Whether the developed scenarios and their results are transferable to other cases requires further validation. The transferability of indicators that are applied across member states (e.g. the farmland bird index) is probably higher than for country specific indicators. However, the improvements achieved in the different scenarios show ways of enhancing data quality and/or quantity which are expected to be useful for varying indicators or methods. A number of lessons can be derived for future environmental monitoring programmes:

- Setting data requirements at the beginning of each programming period facilitates sound statistical analyses of environmental impacts and robust recommendations
- Planning of impact evaluations at the stage of scheme design helps to ensure necessary data availability for consistent evaluation
- Adjustments to sampling and monitoring methods targeted at RDP evaluation can improve cost-effectiveness of the evaluation process
- Embedding additional data collections into a multi-purpose monitoring system eventually leads to resource savings of the public sector and more comprehensive data sets.

7 References

Aalders, I, Morrice, J, Miller, D, Schwarz, G (2015) Report on the theoretical and methodological framework at macro level. Report to the European Commission. Deliverable D5.3, ENVIEVAL project (Project Reference: 312071).

Artell, J., Aakkula, J., Toikkanen, H. (2015) Summary report on the methodological framework for counterfactual development. Report to the European Commission. Deliverable D3.3, ENVIEVAL project (Project Reference: 312071).

Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A. and Vincke, P. (2006) Evaluation and Decision Models with Multiple Criteria: Stepping stones for the analyst. Springer Science and Business Media Inc.: New York.

Carlson M, Schmiegelow F (2002) Cost-effective Sampling Design Applied to Large-scale Monitoring of Boreal Birds. Conservation Ecology 6(2): 11. [online] URL: http://www.consecol.org/vol6/iss2/art11/

DCLG (2009) Multi-criteria analysis: a manual. Department for Communities and Local Government: London.

EC (2001) 144 final, Communication from the Commission to the Council and the European Parliament. Statistical Information needed for Indicators to monitor the Integration of Environmental concerns into the Common Agricultural Policy.

EC (2012) Eurostat Statistics explained. Agri-environmental indicator - population trends of farmland birds. http://ec.europa.eu/eurostat/statistics-explained/index.php/Agri-environmental_indicator_-_population_trends_of_farmland_birds (Accessed: 21st Oct. 2015).

ESA (European Space Agency) (2015). https://earth.esa.int/web/guest/missions/esa-future-missions/sentinel-2. Accessed: 09th of October 2015, 13.43 pm.

Faehrmann, B. and Grajewski, R. (2013) How expensive is the implementation of rural development programmes? Empirical results on implementation costs and their consideration in the evaluation of rural development programmes. European Review of Agricultural Economics, 40, 4, 541-572,

FERA (The Food and Environment Research Agency) (2009) Natural Heritage Outcome Monitoring: Pre-project Scoping Study on Methodology Options. Final Report. 2007-2013 Scotland Rural Development Programme.

Kelemen, E., Podmaniczky, L., Balázs, B., Choisis, J-P, Gomiero. T., Paoletti, M. and Sartho, J-P (2011) Report on the farmers' perception of biodiversity indicators associated to organic and low-input farming systems. Deliverable D4.4, BioBio project.

Lindenmayer DB, Zammit C, Attwood SJ, Burns E, Shepherd CL, Kay G, Wood J (2012) A Novel and Cost-Effective Monitoring Approach for Outcomes in an Australian Biodiversity Conservation Incentive Program. PLoS ONE 7(12): e50872.

LWK Niedersachsen (2014) Berechnungsgrundlage der Landwirtschaftskammer Niedersachsen. Blaubuch Erntejahr 2014.

NLWKN (2015) Anwenderhandbuch für die Zusatzberatung Wasserschutz Grundwasserschutzorientierte Bewirtschaftungsmaßnahmen in der Landwirtschaft und Methoden zu ihrer Erfolgskontrolle. Grundwasser, Band 21.

NLWKN (2010) Untersuchung des mineralischen Stickstoffs im Boden. Empfehlungen zur Nutzung der Herbst-Nmin-Methode für die Erfolgskontrolle und zur Prognose der Sickerwassergüte. Grundwasser, Band 8.

NONIE (2009) Impact Evaluations and Development: NONIE guidance on impact evaluation. NONIE, Washington. http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf

Nijkamp, P. and Blaas, E.W. (1994) Impact Assessment and Evaluation in Transportation Planning. Springer Publishers.

Povellato, A, Lasorella, MV, Longhitano, D (2015) Summary report on the theoretical and methodological framework at micro level. Report to the European Commission. Deliverable D4.3, ENVIEVAL project (Project Reference: 312071).

Saaty, T.L. (1980) “The Analytic Hierarchy Process.” McGraw-Hill, New York

Schmidt TG and Osterburg B (2010) Wirkung von Agrarumweltmaßnahmen auf den mineralischen Stickstoffgehalt von Böden und Kostenwirksamkeit der Maßnahmenumsetzung. WAgriCo 2 Projektbericht.

Targetti, S, Herzog, F, Geijzendorffer, FR, Wolfrum, S, Arndorfer, M, Balàzs, K,, Choisis, J.P., Dennis, P, Eiter, S, Fjellstad, W, J.K. Friedel. J.K. Jeanneret, P, Jongman, R.H.G., Kainz, M., Luescher, G, Moreno, G, Zanetti, T, Sarthoum, J.P. Stoyanova, S., Wiley, D,. Paolettil, M.G , Viaggi, D (2014) Estimating the cost of different strategies for measuring farmland biodiversity: Evidence from a Europe-wide field evaluation. Ecological Indicators, 45, 434–443.

Tulloch A, Possingham HP, Wilson K (2011) Wise selection of an indicator for monitoring the success of management actions. Biological Conservation 144 (2011) 141–154.

Van Delft, A. and Nijkamp, P. (1977) Multi-Criteria Analysis and Regional Decision-Making. Studies in Applied Regional Sciences: Leiden.