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: Development and application of methodologies and tools, including indicators, for the assessment of environmental impacts of rural development programmes in the EU

Report D3.3
Summary report on the methodological framework for counterfactual development

---

Authors: Janne Artell, Jyrki Aakkula, Heini Toikkanen (LUKE)

Approved by Work Package Manager of WP3: **Jyrki Aakkula (LUKE)**

Date: September 2015

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

Date: September 2015

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

List of Acronyms

AEP	Agri-Environment Payments
CAP	Common Agricultural Policy
CGE	Computable General Equilibrium
CMEF	Common Monitoring and Evaluation Framework
CMES	Common Monitoring and Evaluation System
DiD	Difference-in-Difference
FADN	Farm Accountancy Data Network
GNB	Gross Nutrient Balance
HNV	High Nature Value
IACS	Integrated Administration and Control System
IV	Instrumental Variable
NB	Naïve baseline
NG	Naïve group
PM	Pipeline Methods
PSM	Propensity Score Matching
PSM-DiD	Propensity Score Matching Double Difference
QA	Qualitative Analysis
RD	Regression Discontinuity
RDP	Rural Development Programme
SE	Structural Econometric

Executive summary

The main issue in constructing the counterfactual is programme participants self-selecting to the programme. In such a case, a simple comparison between participants and non-participants gives a biased estimate of the environmental effects of the programme.

The logic model for constructing a counterfactual in evaluating the environmental public good impact determines the possible routes of analysis with piece-by-piece evaluation of the available data. The logic model identifies methods for constructing a counterfactual, and explicitly forces the evaluator to assess the possibilities and weaknesses of the available data. There are three general level options for counterfactual construction: evaluation options without comparison groups, qualitative and naïve quantitative evaluation options with an ad-hoc approach to sample selection, and statistics-based evaluation options with an explicit approach to sample selection.

A comparison of experiences from the ENVIEVAL public good evaluation case studies shows that the logic model serves the purpose of identifying issues and methodological options opened by existing data. Further, the experiences show that despite the general level guidance of the logic model, data-related issues, and in some cases problems in the methods themselves, can complicate evaluation. As problems are case specific, so are the solutions. However, this report serves to give examples of different cases in constructing a counterfactual. The report does find, however, that current data can be used for evaluation, and can be improved to allow rigorous evaluation procedures - Data availability, timing of monitoring data and its spatial resolution all need actions.

Generally, the case studies using direct impact indicators (i.e. the actual environmental impact) had more often data availability related problems. Monitoring is expensive, leading to relatively small data repositories which may not cover non-participant farms or areas. Further, actual environmental impacts are often slow processes which may not exhibit during the evaluation period even with good data availability. In these cases modelling approaches, qualitative analysis or using proxy indicators (e.g. indirect pressure indicators) in statistical analysis may be preferable.

The logic model shows promise in finding different approaches for constructing a counterfactual with varying data availability. As some case studies discovered, data quality as initially perceived did not always allow the intended method for constructing a counterfactual, but the logic model could in those cases be used to trace back other methodological options. Additionally the case studies show that sometimes an ad-hoc approach to sample selection may help in improving the evaluation quality from simple comparisons. Ad-hoc approaches rely on the evaluator's ability to understand the underlying selection process to measures. However, such approaches are not easily comparable over time and across measures warranting improved monitoring that specifically aims to help evaluation process.

The structure of the methodological framework for counterfactual development will inform the development of the methodological handbook for the evaluation of environmental impacts of rural development programmes, which will pay particular attention to the practical problems encountered as well as the solutions applied in the environmental public good case studies.

1 Background

1.1 Rationale for the framework

The simplified framework for good evaluation (Figure 1) has the development of a counterfactual as the second step following the generalized logic model. The generalized logic model sets the base for evaluation by first identifying the evaluation question, the available data from the actors participating, and those not participating in the evaluated measure or programme[1], and the environmental indicator data[2] that can be causally linked to the evaluated measure and evaluation questions. The causal link between indicators and impacts sough after in the evaluation questions is a key building block in impact analysis. In the background, a qualitative assessment on external factors affecting the environmental indicator and measure participation is necessary to establish a level of certainty between the causality.

Determining a counterfactual is essential for any evaluation. The counterfactual is effectively a point of comparison, which is typically defined as the state of the world without the evaluated measure. Thus all impact evaluations develop a counterfactual, either implicitly or explicitly. It is the purpose of this deliverable to lay the basis for transparent development of a counterfactual or a number of counterfactuals. The counterfactual layer sets the base for micro[3]- and macro[4]-level evaluations, which further refine the methodological options for the evaluator with the available data.

We note that the logic model can be used in two ways. First, the logic model can be used in ex-post evaluations with fixed data sources to find a set of possible methods to construct the counterfactual. Second, the logic model can assist in the planning of data collection routines for future evaluations by going in reverse order from the methods up to data requirements.

[1] In later text we refer only to *measure* for simplicity, but return to the usability of the logic model in the evaluation of programmes containing multiple measures in section 3.1 Generally on the applicability of the counterfactual logic model.
[2] This includes common, additional and programme specific result and impact indicators.
[3] For more detail see Povellato et al. (2015) on the theoretical and methodological framework for micro-level.
[4] For more detail see Aalders et al. (2015) on the theoretical and methodological framework for macro-level.

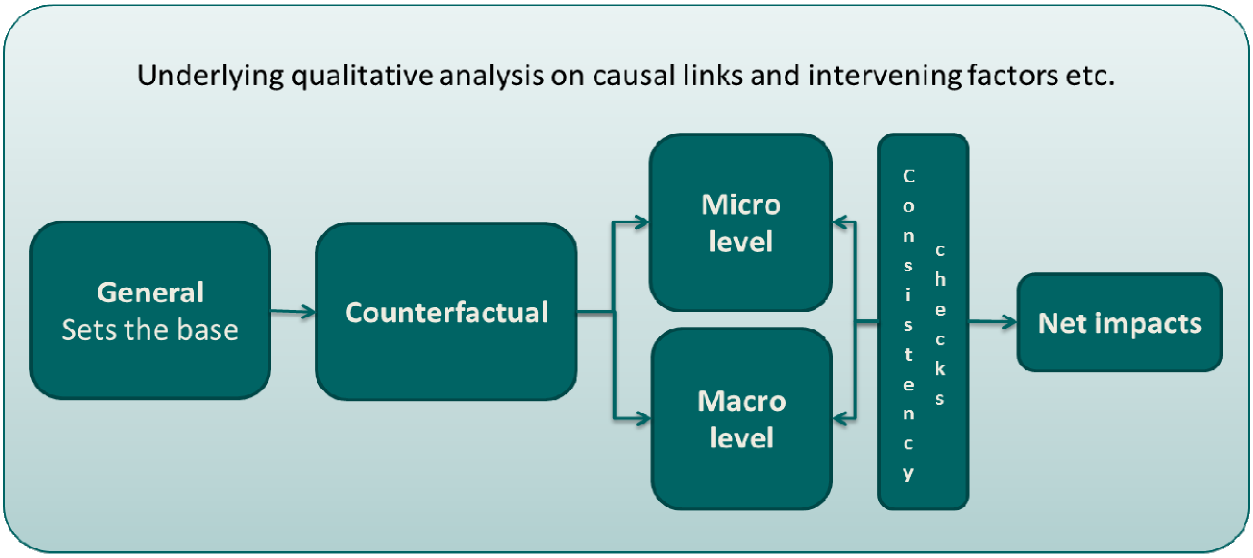

Figure 1 Simplified logic model flow of evaluation

1.2 Analysis using a counterfactual

A counterfactual is not a method per se. It is a point of comparison, the most likely state of the world without the evaluated measure. All evaluations make either an implicit or explicit assumption of the counterfactual since we cannot observe something that has not occurred, only the historical path leading to the present.

There are a number of methods to construct a counterfactual[5]. These methods can be broadly categorised into three groups by their data requirements and the how they handle problems related to measure self-selection (i.e. sample selection, see next section for further discussion). In a decreasing data requirement order, the groups and the underlying methods to construct a counterfactual are:

i. Statistics-based Evaluation Options – Explicit Approach to Sample Selection (heavy data requirement)
   a. Difference-in-differences[6] (DiD)
   b. Propensity score matching (PSM)
   c. Joint DiD and PSM (DiD-PSM)
   d. Other advanced regression techniques, including but not limited to
      i. Instrumental variables methods (IV)
      ii. Regression discontinuity methods (RD)
      iii. Pipeline methods (PM)

[5] The methods we list are used when the treatment (i.e. the evaluated measure) is not randomly distributed across farmers. The participation is up to the farmers to decide, making measure-non-participants likely to differ from participants. There are also special cases where the programme participation can be considered a quasi-experiment, i.e. the programme participation status mimics random participation. We direct the reader to ENVIEVAL Report D3.1 Review of counterfactual methods (Artell et al., 2013) for further discussion.

[6] May also fall under naïve treatment of sample selection in case the underlying trends in the evaluated indicator across measure participants and non-participants are separate and not taken into account in the evaluation.

ii. Evaluation options without comparison groups (low to heavy data requirement)
    a. Advanced environmental economic models, including but not limited to
        i. Computable general equilibrium sectoral economic models (CGE)
        ii. Structural econometric models (SE)
        iii. Bio-economic modelling approaches (BE)

iii. Qualitative and naïve quantitative evaluation options – Ad-hoc approach to sample selection (low data requirement)
    a. Qualitative analysis (QA)
    b. Naïve baseline comparison (NB)
    c. Naïve group comparison (NG)
    d. Naïve combined NB & NG comparisons

1.3 Evaluation challenges

Artell et al. (2013) identified a number of challenges in practical evaluation situations in their review of counterfactual methods:

- There are too many (environmental impact) indicators for national administrations to handle
- The focus is too much on indicators of outputs rather than outcomes
- Causality between measures and impacts increases in vagueness with multiple intervening factors (e.g. CAP and regional policies)
- Identification of impacts is complex due to different spatial scales of environmental outcomes and RDP implementation regions
- There are gaps in data and existing data suffers from errors
- Data may be difficult to obtain and use due to rigid data storage choices and other restrictions
- Evaluation capacity within Member States is lacking.

Before a counterfactual is developed, the causal link between the evaluation question environmental indicator and the evaluated measure must be established. Particularly for methods constructing the counterfactual through modelling, the causal link must be expressed in a functional form in mathematical terms. For other methods, the environmental indicator must be defined in some quantifiable metric or in distinct categories. The environmental impact may also be hard to attribute to a single actor (micro-level evaluations) implementing the measure due to diffuse pollution, wide-spread environmental impact, or long-term environmental response to measures. In cases like diffuse water pollution or climate change effects, the choice of environmental indicator is usually shifted from the actual impact to pressure indicators, e.g. changes in gross nutrient balance or greenhouse gas emissions. Alternatively, the focus can be shifted to macro-level evaluation, including larger spatial areas. This approach often suffers from multiple intervening factors that also affect the development of the environmental indicator over time.

In most cases the evaluation challenges are due to either data quantity or quality. When external factors affect the environmental indicator development over time or measure participation, data related issues in evaluation typically worsen and require at least a qualitative assessment on the severity and direction of effect from the intervening factors.

As stated before, the counterfactual is a point of comparison. The point of comparison may be constructed statistically from actual data, using models that exploit actual data, or a heuristic approach. A challenge that motivates the use of statistical or (bio)economic modelling approaches is that measures are not implemented randomly. If measures were implemented randomly across farms, a comparison of the environmental indicator between participants and non-participants to the measure would suffice for evaluation. However, as farmers are likely to self-select to participate in measures that best suit them, a naïve comparison between participants and non-participants is likely to cause a biased estimate of environmental impact[7]. The evaluators may run into situations where data exists only for participants.

Other complications arise with multiple types of participation during the programme period: those who always or never participate, those who drop out or join late from the measures, and those who have previously participated in the same or similar measure, where the impacts carry through to the evaluation period. It may be difficult to acquire a sufficient amount of data for a statistical evaluation taking account all groups, leading to a counterfactual that may not describe a world without the measure. A coherent programme-level evaluation requires the counterfactuals used to evaluate separate measures – if the evaluation is conducted piece-by-piece – to be the same.

2 Counterfactual development logic model

2.1 Generally on the applicability of the counterfactual logic model

The following steps describe the workflow in going through the logic model to design a counterfactual. The structure is general and works for both micro and macro stages, providing a set of questions that need to be answered before deciding on the counterfactual design.

The logic model for developing a counterfactual focuses on data availability and quality challenges as they drive the possibilities for using different methods for evaluation. Underlying the model is the requirement for the environmental indicator to have a causal link to the evaluated measure. The requirement highlights the difference between measure specific and more general programme level impact evaluation.

Data requirements increase in tandem with complexity of the evaluation question. Since the methods for constructing a counterfactual are data sensitive, a general level evaluation question may force the use of equally general methods. By setting the evaluation question first at a disaggregated measure level and then assessing the possible level of analysis with existing data using the logic model, it is possible to identify the critical data gaps in more general programme-level evaluation. In addition, this approach reveals if the data allows similar counterfactuals to be constructed for each measure.

Further, the logic model works best for cases with a single indicator. This extends to cases where a measure or a group of measures aim for the same environmental outcome. Programme-level evaluations are often too abstract to be evaluated with a single indicator and, hence, a single evaluation method. Rather, the logic model offers tools with which to construct a sound

[7] The sample selection issue is further discussed in the ENVIEVAL Report D3.1 Review of counterfactual methods (Artell et al., 2013).

counterfactual-based evaluation that may include a number of approaches in programme-level impact evaluation[8]. Thus for a programme level evaluation, the evaluators should use the logic model to collect and compare the different evaluation possibilities and generate a concise evaluation plan that uses similar counterfactual scenarios across the line. In practice, the last requirement can be challenging due to data gaps. The logic model, however, offers a structured identification plan for counterfactual scenarios enabling discussion on the impact of different counterfactual scenarios to the programme-level evaluation

We note that due to the number of different evaluation methods and the need to keep the logic model as uncomplicated as possible, the final suitability of the available data for each method, especially for the statistical methods, must always be assessed case-by-case with experts. It is possible that the logic model suggests a statistical approach that cannot be conducted with the data. Each comparison group in statistical analysis should have at least 30 observations of the functional units. Further, we have assumed in this logic model that the data are spatially and temporally synchronized with the functional units and the programme period. In cases where this assumption does not hold, the evaluator needs to qualitatively assess the risks and magnitude of bias in the evaluation results.

The logic model drives for quantitative approaches over qualitative. However, careful qualitative assessment should underlie each stage of the logic model[9], e.g. the identification and choices of suitable comparison groups for evaluation are subject to qualitative approaches the more complex the setting.

2.2 The logic model

In the following text we will go through the logic model part-by-part with explanations and real-world examples from the ENVIEVAL case studies. Figure 2 shows the full logic model for designing counterfactuals.

[8] If the evaluation uses multiple indicators, statistical approaches can provide comparable counterfactuals at the extent of the indicator with least data. Modelled approaches for counterfactuals allow for multiple indicators so long as there are mathematical representations of the indicators. Qualitative and naïve approaches are the least restrictive for using multiple indicators, as the counterfactual is based on a more heuristic approach.

[9] This qualitative assessment is important especially in cases where the intervention logic of the measure or policy and data collection have not been well linked in practice.

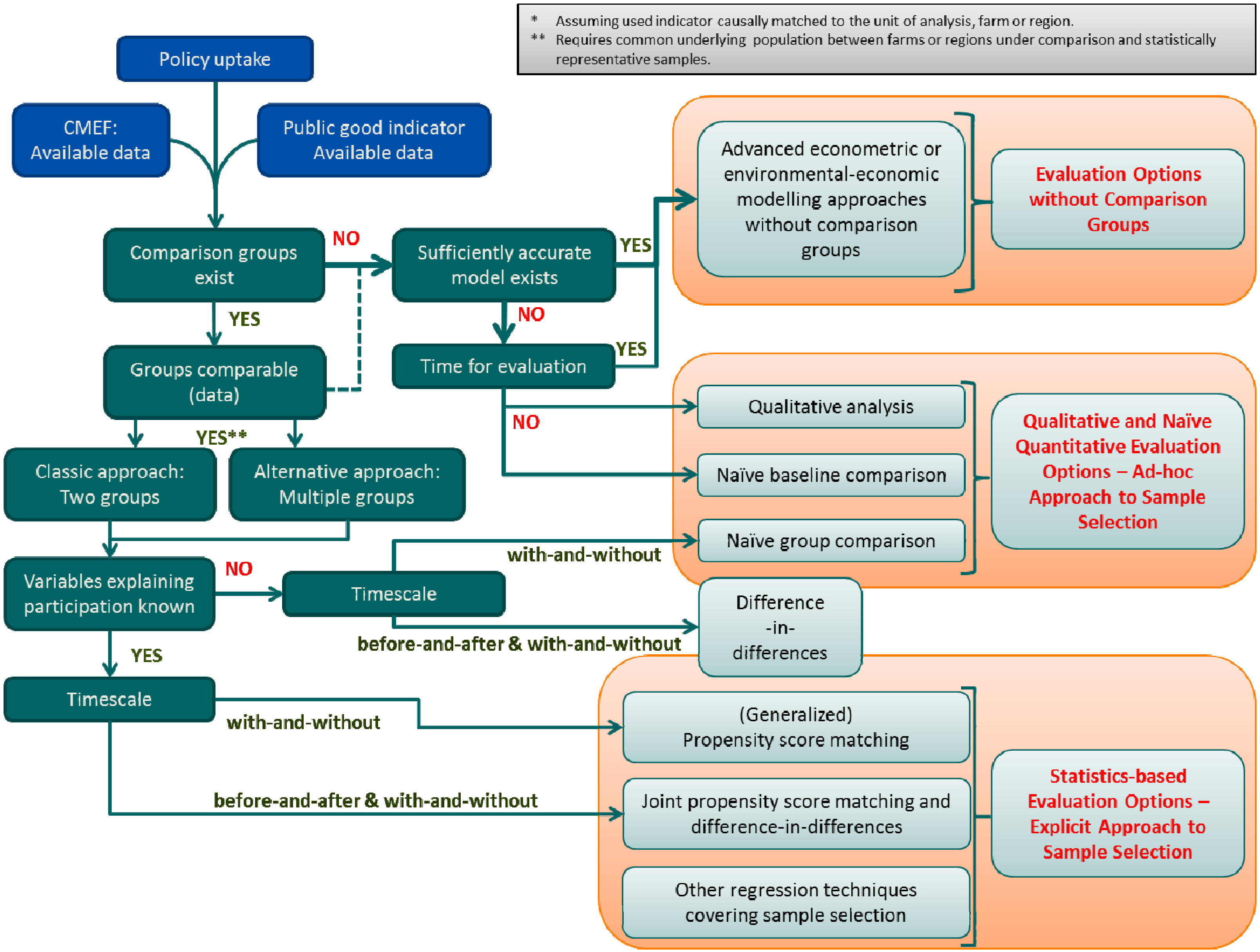

Figure 2 Workflow and description of the counterfactual design

2.2.1 Step 1: Compiling multiple data to a single dataset

The initial stage of the counterfactual logic model (Figure 3) compiles the available data for evaluation. The Common Monitoring Evaluation System (CMES) and Framework (CMEF) provides data for the official common evaluation question(s) and indicators, whereas public good indicators are case specific, when the CMEF indicators do not suffice or are inferior in measuring the environmental impact at the required level[10]. Policy uptake data determines the possible actual comparison groups in later stages, initially indicating the farms that have partaken in the evaluated measure and those that have not. At this stage the logic model is not strict on the micro or macro level of analysis - the methods of comparison do not largely disagree with scale. However, the micro and macro specific logic models further refine the possible approaches to counterfactuals.

The evaluator should confirm that the data from each source can be (causally) linked to a common functional unit, e.g. farm-level or some aggregate level spatial unit.

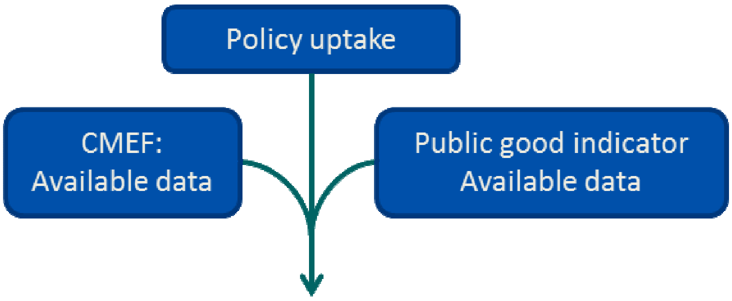

Figure 3 Input to the counterfactual logic model

2.2.2 Step 2 Determining comparison groups

The next step is to determine whether the comparison groups exist at all to construct a counterfactual. For example, in the Finnish case studies, the participation rate to agri-environment payments (AEP) is so high that a comparison group (non-participants or other partial participation groups) does not effectively exist[11]. In such a case, the logic model directs the evaluator to Step 3 for other than statistically-based counterfactuals.

If comparison groups should exist, the next step is to assess if data on both participants and non-participants are available for evaluation. Note that non-participants are not the only possible comparison group usable for constructing a counterfactual. However, they are the only source of data that allow for constructing a statistics-based counterfactual for a state of the world *without* the evaluated measure. Other comparison groups allow for intermediate impact estimates, for example, when information exists for partial participation in the evaluated measure.

While the logic model considers comparison groups predominantly from a data perspective, their identification is important even when data does not allow for their full individualization. The recognized comparison groups work as a basis for deciding the final point of comparison, i.e. the

[10] Member states are expected to define additional and programme specific result and impact indicators (as well as programme specific evaluation questions). The case studies provide evidence on the suitability of the tested non-CMES indicators as additional or programme specific result and impact indicators.

[11] The non-participant population differs too much from the participant population to allow for a meaningful comparison.

counterfactual, in cases where the logic model guides to evaluation approaches that are less dependent on statistics.

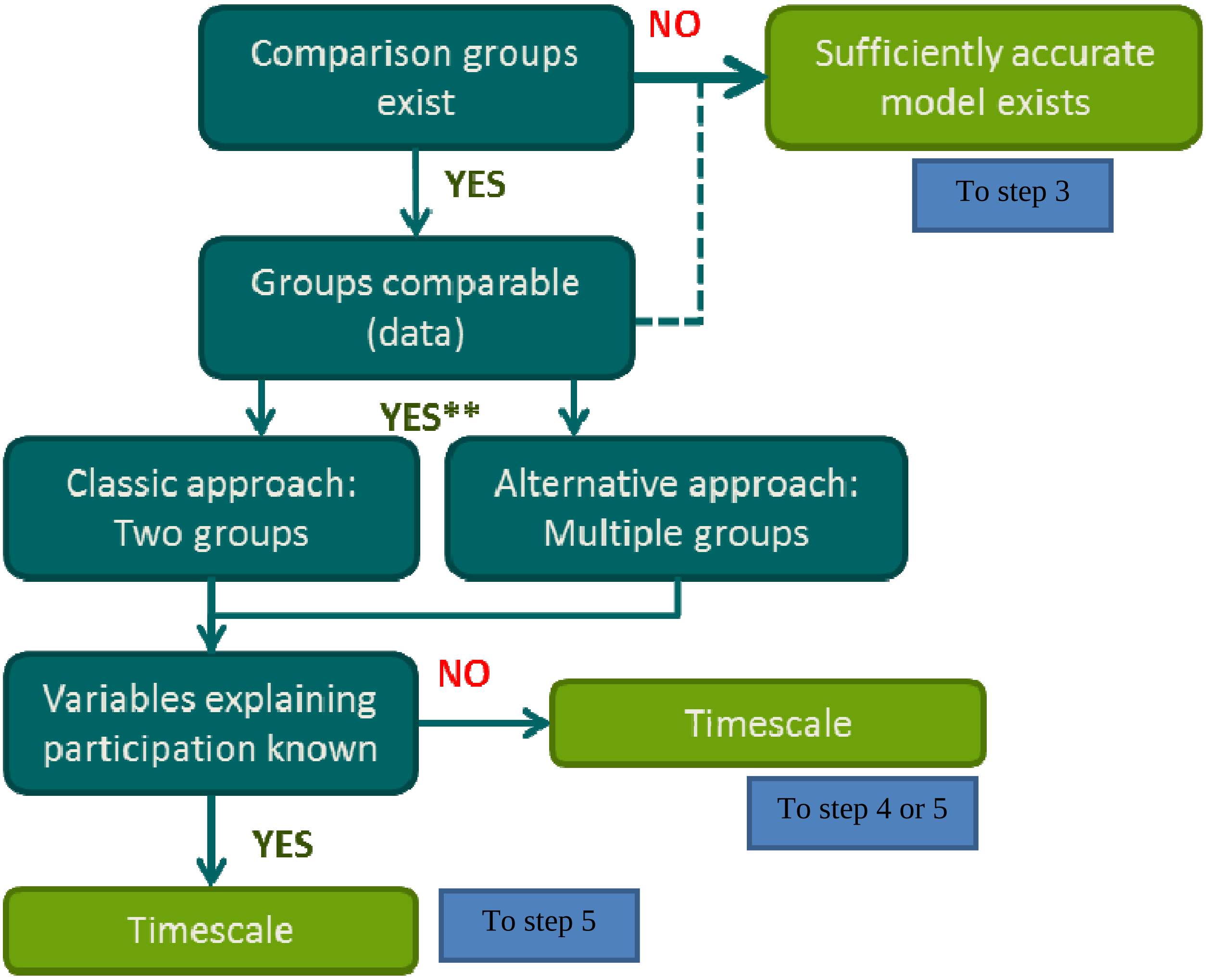

Figure 4 Comparison group definition stage, step 2

Table 1 elaborates on the different possible comparison groups that can be identified from data[12]. In addition to participation status, comparison groups can arise from significant intervening factors. Such factors are, for example, internal inertia of environmental effects due to participating measures during prior programme periods, and external inertia in the form of current or historical environmental pressure affecting the indicator. The table is not exhaustive, as intervening effects are many in form. However, the table serves to show the increasing complexity for statistical approaches with multiple effects. Qualitative approaches may be needed to describe the possible severity of each effect and narrow down the comparison groups to a reasonable number.

[12] The ENVIEVAL D3.2 Report on monitoring and data requirements for counterfactual methods (Artell, 2014) explains the comparison group formation in more detail.

Table 1 Comparison group formation with intervening internal and external factors

Participation status in evaluation period	Eligibility rules exist for participation	Internal inertia	External inertia	Minimum number of groups
Only participants/ non-participants (2)	All eligible (x1)	None (x1)	None (x1)	2
			Historically significant outside pressure at min. one area (+2)	4
		Previous participation status affects environmental effects or participation probability (x2)	None (x1)	4
			Historically significant outside pressure at min. one area (+2)	6
	Some non-participants ineligible or in a queue to participate (+1)	None (x1)	None (x1)	3
			Historically significant outside pressure at min. one area (+2)	5
		Previous participation status affects environmental effects or participation probability (x2)	None (x1)	6
			Historically significant outside pressure at min. one area (+2)	8
Participants/ non-participants, also drop outs and/or late joiners (3/4)	All eligible (x1)	None (x1)	None (x1)	3/4
			Historically significant outside pressure at min. one area (+2)	5/6
		Previous participation status affects environmental effects or participation probability (x2)	None (x1)	6/8
			Historically significant outside pressure at min. one area (+2)	8/10

Participation status in evaluation period	Eligibility rules exist for participation	Internal inertia	External inertia	Minimum number of groups
	Some non-participants ineligible or in a queue to participate (+1)	None (x1)	None (x1)	4/5
			Historically significant outside pressure at min. one area (+2)	6/7
		Previous participation status affects environmental effects or participation probability (x2)	None (x1)	8/10
			Historically significant outside pressure at min. one area (+2)	10/12
No non-participants (1)	All eligible (x1)	None (x1)	None (x1)	no statistical comparison possible
		Previous participation status affects environmental effects (x2)	None (x1)	2, no counterfactual for inaction in statistical comparison
			Historically significant outside pressure at min. one area (+2)	4, no counterfactual for inaction in statistical comparison
No non-participants but late joiners (2)	All eligible, queues to participate (x1)	None (x1)	None (x1)	2, counterfactual for partial measure participation only in statistical comparison
			Historically significant outside pressure at min. one area (+2)	4, counterfactual for partial measure participation only in statistical comparison
		Previous participation	None (x1)	4, counterfactual for partial measure

Participation status in evaluation period	Eligibility rules exist for participation	Internal inertia	External inertia	Minimum number of groups
		status affects environmental effects (x2)		participation only in statistical comparison
			Historically significant outside pressure at min. one area (+2)	6, counterfactual for partial measure participation only in statistical comparison

As statistical methods have different routines for more than two comparison groups, a parallel stage is identified for analysis with multiple comparison groups. The usability of each identified comparison group[13] in statistical analysis is dependent on the quality and quantity of data. For example, in the Greek diffuse water pollution case study multiple comparison groups existed, but access to temporal data was limited. The restricted data did not allow a statistical comparison of multiple groups, thus shifting the choice of counterfactual development to a simple with-and-without group comparison (step 4).

After the number of comparison groups supported by data has been identified, the evaluator should examine if the data includes variables that help to explain why different comparison groups have selected their participation status. These variables address sample selection bias in statistical analysis, where comparison groups may differ by population type due to different underlying qualities.

Variables that explain participation to the evaluated measure or policy are case-specific and depend on the functional unit (e.g. regional vs. farm uptake of a measure). These variables should, at the very least, include factors of functional units that are targeted in the measure or policy. For example, if the policy targets cereal producers below certain income level in high risk erosion areas, at least farmer income, production type and having fields in high risk erosion areas should be known for all comparison groups. Farm-level data is best suited to identify where measures have been taken, and which farms have participated. For regional or coarse spatial resolution data, typical for macro-level analysis, this level of analysis may be difficult[14] or impossible, and the evaluator may be forced to exit at step 4. Provided data is abundant for all comparison groups, the evaluator may take the exit at step 5.

[13] Comparison groups may also include composite groupings of similar groups to prevent the data to fragment too much for sound analysis.

[14] Some spatially aggregated area data may pass this step if the level of participation can be identified as a percentage and the data is otherwise sufficient and causal links between measures and environmental effects can be made. However, identifying the rationale of those who participate (i.e. reasons for self-selection) in the measure and external intervening factors become rapidly an issue, as the scale of aggregation increases.

2.2.3 Step 3 Evaluation options without (data-based) comparison groups

Stemming from steps 1 and 2, if the data does not sufficiently describe the different comparison groups[15], it should next be assessed if there are models (Figure 5) that could construct the missing comparison groups, i.e. develop the counterfactual. If a model exists that can use the available data to evaluate a state of the world without the evaluated measure at the desired resolution (e.g. representative farm-level resolution or regional resolution), then we suggest using the existing model. In the Finnish case studies for climate stability and diffuse water pollution existing models were tested to describe each case. The climate stability case study employed a sectoral economic model to describe in a regional level the effects of AEP to agricultural production and thus greenhouse gas emissions. On the other hand, the diffuse water quality case study employed a structural econometric model to construct a response function for a representative farm to assess how farms would on average behave in the absence of AEP.

While existing models offer possibilities for quick, repeatable, and even ex-ante evaluation, the models require accommodation to new data and may be rigid if changes to their structure are needed. A warning example is provided by the Finnish diffuse water quality case, where the model initially produced implausible results using updated data. Recalibrating the model was not possible within the timeframe and resources allotted for evaluation. Nevertheless, we do recommend exploring the option to build a model that enables consistent counterfactual study of effects also in future evaluations, if time and resources allow[16]. Such an approach enables a self-updating system for evaluation with new data and possibilities to build on earlier modelling work. It should be noted that the evaluator may also choose this option over statistical approaches if the model approach is the cost-efficient solution.

An important note in using modelled counterfactuals is that the evaluator has full control and responsibility in choosing the point of comparison. A number of scenarios are possible in modelling, including a simple business-as-usual scenario without the evaluated measure. Such counterfactual scenarios could be decided to follow the most credible path should the measure not have been implemented. The evaluator needs to decide if other intervening policies or measures should be taken into account in the counterfactual scenario.

Where there are no resources to build a model, the evaluator is guided to step 4.

[15] Especially the non-participant group, when the counterfactual is required to depict a world without the evaluated measure.

[16] Modelling approaches provide a good evaluation tool especially for widely implemented horizontal policy measures.

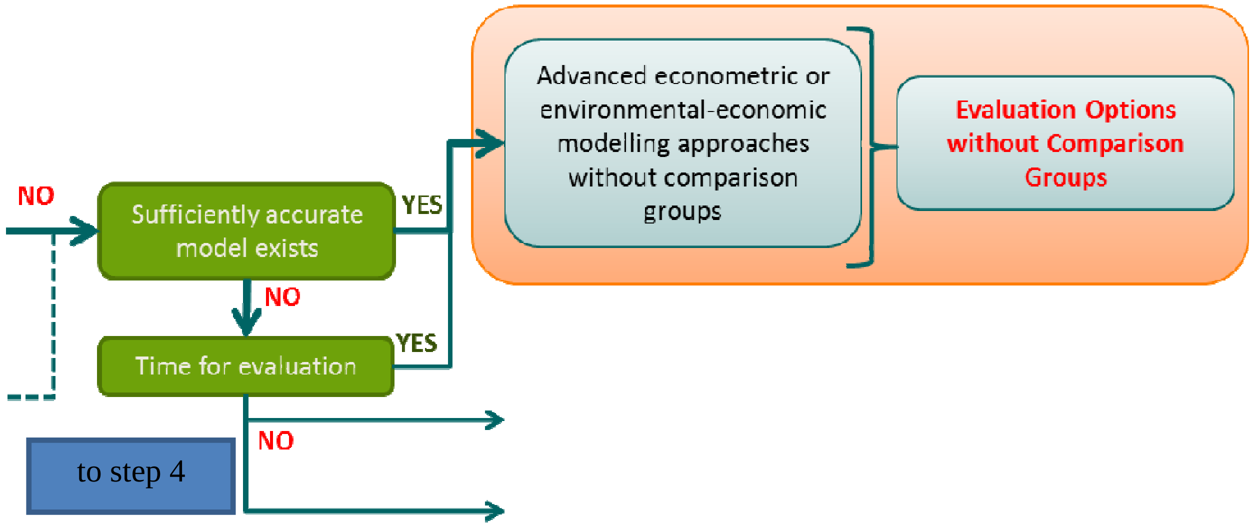

Figure 5 Choice of long run evaluation options without comparison groups for counterfactual construction, step 3

2.2.4 Step 4 Qualitative and naïve quantitative evaluation options – Ad-hoc approach to sample selection

The evaluator can arrive at step 4 via multiple routes, i.e. through steps 2 and 3. Essentially, step 4 considers evaluation options where data and existing models are insufficient for elaborate quantitative evaluation options.

Coming from step 2, if the data on comparison groups does not include additional information, i.e. variables, that would explain why farms would participate or self-select to the evaluated measure, the next step is to determine if the data covers multiple points in time. If there is sufficient data for participants and non-participants for both before and after the evaluation period, the evaluator can resort to the difference-in-difference (DiD) method family. The DiD can be considered either a naïve quantitative (step 4) or a statistics-based (step5) approach depending on the underlying assumption on data quality. If the evaluator believes the participant and non-participant groups would have different time trends for the environmental indicator development, the evaluator can opt for an ad-hoc sample selection correction in the data. For example, in the Italian climate stability case study, the evaluator chose the sample for analysis based on experience to compare neighbouring participants and non-participants. In the absence of dissimilar time trends across comparison groups, the DiD is no longer considered as a naïve quantitative approach to sample selection.

When data is available only for cross-section at the end of the evaluation period, the evaluator has to conduct naïve quantitative analysis comparing the environmental indicator level with different comparison groups. The typical counterfactual would be provided the comparison group that has not participated in the measure. Following the naïve estimation of the effects, the evaluator should make a careful qualitative assessment on what internal and external factors are likely to affect the evaluation results. Is it likely that if there was self-selection of the measure, are the participants similar to non-participants in other respects? If there is self-selection, is the bias in the evaluation results likely an over- or underestimate, and how severe is the bias likely to be?

Coming from step 3, if there is no existing model, or time to make one, that could construct the counterfactual, the evaluator cannot make a fully data-based assumption of the counterfactual. In case of a missing comparison group but with some level of data for participants, the evaluator may make a naïve baseline assumption based on qualitative analysis. For example, expert opinions can be used to determine if a trend exists in the development of the environmental indicator. The counterfactual can be based on this trend and then contrasted to the observed level of the indicator. This way the evaluator essentially decides the level of environmental impact from the measure. Therefore the decision, its arguments and how it affects the counterfactual should be well and explicitly documented with qualitative sensitivity analysis if possible.

If the evaluator cannot use data for quantitative analysis for any reason, e.g. lack of data, poor quality of data, poor causal link between the measure and the impact, the only way to evaluate the impact of the measure is through qualitative analysis. There are numerous methods in the qualitative analysis literature to develop a counterfactual, but they often rely on different constructs by expert panels. Qualitative analysis requires careful discussion in the evaluation report showing the thought structure behind the analysis. Care should be taken to keep the results understandable. It is clear that qualitative analysis has a very strong foothold as a part of the evaluation procedure in step 4 as quantitative data cannot be fully used to back the counterfactual development.

Evaluators have been often forced to resort to step 4 type methods due to lack of sufficient data. In future evaluations, if data collection is designed from the beginning to support certain methods, fully-fledged step 3 or step 5 level analyses are possible. Ideally, qualitative approaches would then be integrated in mixed method approaches in combination with advanced quasi-experimental methods (step 5).

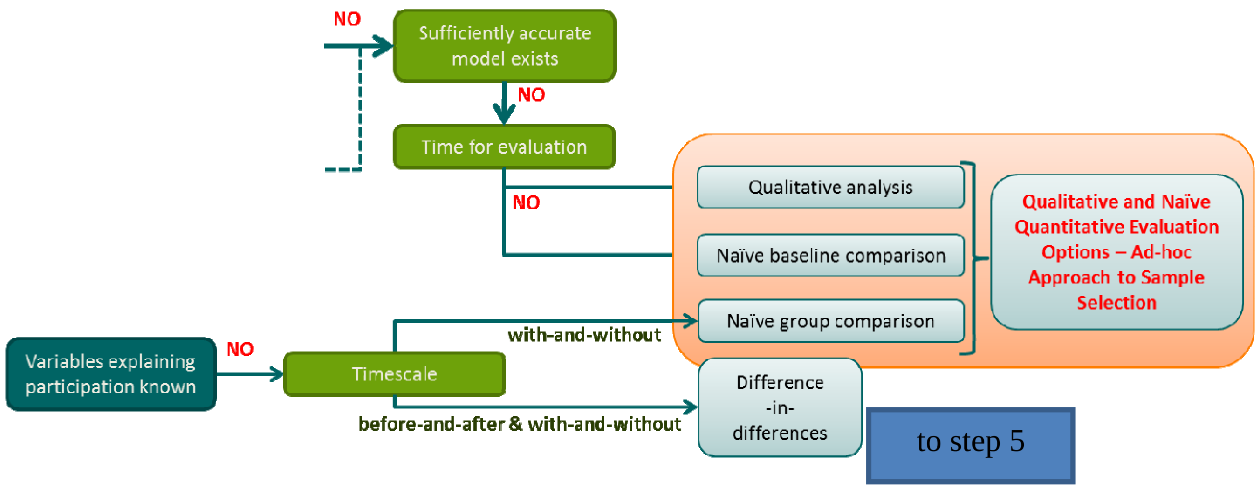

Figure 6 Choice of naïve statistical comparisons or qualitative analysis of counterfactuals, step 4

2.2.5 Step 5 Statistics-based evaluation options – Explicit approach to sample selection

In steps 3 and 4 the evaluator takes a large responsibility in determining the comparison group, whereas, data allowing, in step 5 statistical methods are used to establish a counterfactual that take into account the differences between measure participants and non-participants, i.e. sample selection.

Coming from step 2 there are two critical stages that determine whether statistics-based evaluation options can be used. First, variables explaining the participation in the measure should be known for all comparison groups evaluated. At a minimum the data should cover those participating and those not participating in the measure for the whole evaluation period. The variables explaining participation to the measure are case specific, but include typically the type of produce, size of production, and factors that are targeted by the measure.

Further, the temporal scale of the data should be examined. If data covers the moment before the implementation of the evaluated measure and the end of the evaluation period, the evaluator can resort to statistical methods even if variables explaining participation are unknown. The approach in such a case would be a difference-in-differences comparison of the changes in the environmental indicator during the evaluation period between participant and non-participant groups. In such a case the evaluator assumes that the trend in the environmental indicator is the same between the two groups.

If the evaluator can identify the variables explaining participation, but data[17] exists only for the time after the evaluation period, the evaluator can use the propensity score matching method to construct the counterfactual. The propensity score matching method compares participant and non-participant farms that have similar propensity to participate in the measure, thus alleviating sample selection bias of a naïve group comparison. There must be enough data[18] on similar farms in both groups for the statistical model to work. If the participants and non-participants come from highly different populations, their comparison may not make sense using purely statistical methods. The generalized propensity score matching method can accommodate more than two comparison groups[19], but can be more challenging for the evaluator in terms of data requirements and methodological expertise.

The best opportunities for data-based statistical approaches in developing the counterfactual arise when the data allows a comparison of participant and non-participant groups (i.e. with-and-without) and the situation before-and-after the evaluation period. This is possible in cases where large scale farm monitoring data can be joined with equally precise environmental monitoring data. A clear candidate method of analysis is the joint propensity score matching and difference-in-difference method that enjoys the benefits of both methodologies.

We do note that the statistical and econometric toolbox offers a number of other approaches that tackle sample selection issues. These methods are explored to some extent in the ENVIEVAL Report D3.1 Review of counterfactual methods. In many cases these statistical methods are tailored to the case, and require statistical and econometric expertise from the evaluator. However, if the evaluation question and data remain uniform over time, repeating the evaluation becomes easier with experience. In addition, qualitative analyses and peer-review processes of elaborate statistical evaluations are possible for validation purposes.

[17] The propensity score matching method does, however, greatly benefit from having data on the farms before the programme period on variables explaining participation, as participation itself may cause these variables to change over the programme period.

[18] It depends on the case whether there is enough data for comparison. A hundred observations per comparison group may suffice, where fewer observations likely cause problems in the estimation stage. In the literature, Pufahl and Weiss (2009) used over 10 000 observations, where the two comparison group sizes were roughly of similar magnitude. Notably, Pufahl and Weiss left out a group that participated for some years only in the programme from their analysis. Since participating farms differed significantly from non-participating farms, the application of propensity score matching reduced the number of observations for which matches could be found to approx. 1800.

[19] and participation (treatment) levels representable by a continuous variable.

It is possible to use the elaborate statistical methods for any two (or more) comparison groups. If a clear non-participant group data would not suffice for statistical analysis, but a reasonably close – an almost non-participant – group is identified and has the required data, it can be used to develop a counterfactual that resembles non-participation.

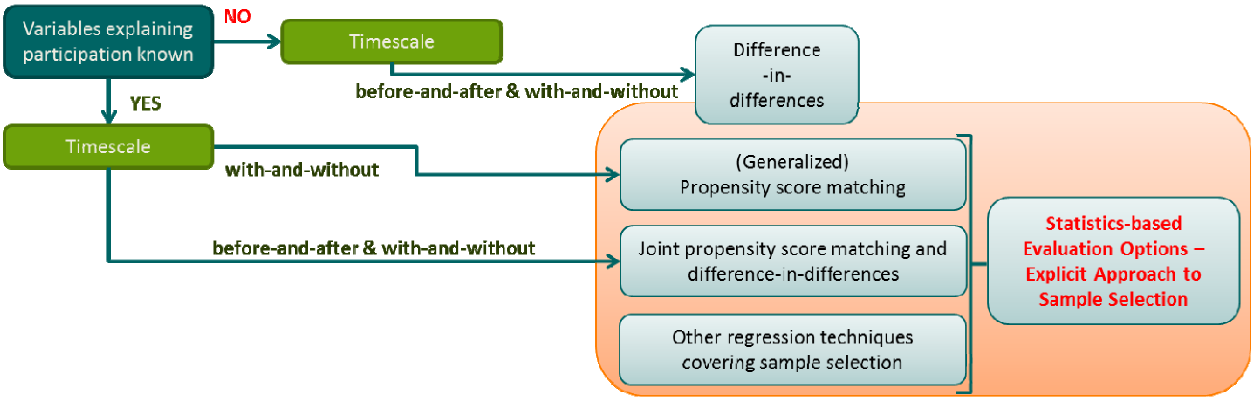

Figure 7 Choice of elaborate statistics-based evaluation for counterfactual construction

2.3 Consistency check considerations

Up- or downscaling evaluation results assume similar conditions at both micro and macro levels ignoring different indirect (e.g. leverage and displacement) effects. Thus it should be qualitatively estimated, whether the conditions affecting the environmental and programme participation selection are similar enough to warrant up- or downscaling. In other words, micro level evaluations with extraordinary conditions should not be upscaled to macro level, and vice versa, without careful consideration of the consequences of such a decision.

3 Synthesis of the logic model development

3.1 Synthesis of the experiences

This section synthesizes the experiences employing the logic model in case studies. The section is divided according to the three general-level counterfactual construction options, i.e. the evaluation options without comparison groups, qualitative and naïve quantitative estimates of counterfactuals with lacking data, and elaborate statistics-based evaluation options. We first go through a list of strengths and weaknesses identified in the earlier work of ENVIEVAL, namely *Report D3.1 Review of counterfactual methods* and *Report D3.2 Report on monitoring and data requirements for counterfactual methods*, followed by observations and experiences found in case studies related.

3.1.1 Evaluation options without comparison groups

Strengths of the approach

- Possibilities to control for sample selection and unobserved effects
- Can accommodate programme and measure level evaluation
- With bioeconomic models environmental outcomes can be estimated with pressure indicator monitoring data
- Ex-ante analysis possible

- Partly evidence-based evaluation with quantitative results
- Regionally explicit spatial analysis possible

Weaknesses of the approach

- Existing models can be rigid in adapting to new evaluation questions
- Construction of a complex model is time- and expert intensive work
- Real-world data for validation may not exist

Problems encountered in case studies

Diffuse water pollution, Finland:

1. Fitting new data required calibration of the model. Calibration proved too time intensive to give results in time
2. Restricted data access due to administrative reasons

Solutions applied to overcome problems in case studies

Diffuse water pollution, Finland:

- 1 and 2: Use results from the model with older data. The older data results have been published in a respected peer reviewed journal giving trust in the robustness of the results. The causal links (transfer functions) between nutrient application, run-off and environmental effects are based on peer-reviewed articles.

Solutions suggested to overcome problems in case studies

Diffuse water pollution, Finland:

- 1. Prepare more time and funding for calibration work
- 2. Discuss with high level officials to overcome administrative restrictions in using FADN data

Synthesis

The Finnish case studies were the only ones employing evaluation options without comparison groups. The climate stability case study suffered from no problems in its application, as the model is continuously used and data updated to the system. However, in the case of the diffuse water pollution case study, two types of problems surfaced. First, acquiring new FADN data was not possible due to administrational obstacles, and second the calibration of the model proved more time intensive than initially thought. Learning from these experiences, while the logic model suggests the use of existing models as a time-saving option especially when no comparison groups can be constructed using data, the evaluator should still take enough time for evaluation, should there be a need to recalibrate the model due to poor validation of results. This raises the importance of reporting the validation process transparently in evaluation. In addition, the evaluator data access should be addressed. Data administration and bureaucracy should not hinder the possibilities of using up-to-date data for evaluation purposes.

3.1.2 Qualitative and naïve quantitative evaluation options – Ad-hoc approach to sample selection

Strengths of the approach

1. Evaluation is less data intensive compared to other methods
2. Addressing causal linkages not limited by data
3. Can accommodate programme and measure level evaluation

Weaknesses of the approach

1. Qualitative results are not necessarily exact
2. Naïve approaches may suffer from sample selection bias
3. Assumptions on baseline development of the indicator cause bias of unknown magnitude
4. Farm-level assessment is not feasible

Problems encountered in case studies

Climate stability, Italy:

- 1. The carbon footprint requires detailed data on matter and energy flows. Complexity induced by the comparison of different farm typologies increases the monitoring costs and restrains sound statistical analysis.
- 2. While data for the carbon footprint had good availability and participant/non-participant sample design, the data exists for only for one point in time and had difficult access due to administrative reasons and poor data storage structure.

High Nature Value areas, Italy:

- 3. The HNV score at farm level lacks data on semi-natural features and number of participant farm observations insufficient for advanced statistical analysis.
- 4. Regional level HNV maps are not updated regularly with sufficiently detailed information in terms of extent and resolution (grain), linking participation and environmental monitoring data is difficult.

High Nature Value areas, Lithuania:

- 5. Land use and forestry cadaster data was either unreliable or the spatial resolution of data was insufficient for quality assessment.

Diffuse water pollution, Greece:

- 6. IACS georeferenced data for participants and non-participants accessible only for one year - no elaborate statistical approach could be applied as farm level participation data was also missing.

Diffuse water pollution, Germany (Nmin indicator):

- 7. We expected to receive micro-level data for recent years. However, only aggregated data was received restraining sound statistical analysis for comparison groups.
- 8. Lack of information on farm structure and management data did not enable the application of advanced matching techniques to improve the net impact assessment.

Soil quality, Scotland (Soil carbon):

- 9. The currently available data on soils do not support an assessment of the impact of RDP measures on soil quality and did not allow the combined use of PSM and DiD and sample selection issues could not be addressed in a robust way.

Animal welfare, Germany (animal-based indicators):

- 10. Availability of animal welfare monitoring data restricts the use of animal-based indicators to counterfactuals and comparison groups with an ad-hoc approach to sample selection issues. The application of the indicators depends on available monitoring data, which impacts on the practical applicability in particular in short term evaluations.

Solutions applied to overcome problems in case studies

High Nature Value areas, Lithuania:

- 5. The case study was a spatial assessment of the extent of HNV land.

Diffuse water pollution, Greece (GNB and water/ha indicators):

- 6. A naïve participant and non-participant comparison using the IACS georeferenced data of 2011. The land parcel level was used as the participation decision level. A farm-level analysis is going to be applied later. The evaluated scheme had simple structure, hence mitigating the difficulties.

Diffuse water pollution, Germany (Nmin indicator):

- 7. Older data were used to apply a matching approach.

Soil quality, Scotland (Soil carbon):

- 9. The assessment is reliant on modelled data for the CMEF indicator. The modelled data represent a sub-catchment level which is the most detailed level feasible for the modelling methodology. Checks of differences in means were carried out for the two comparison groups in before and after situations.

Animal welfare, Germany (animal-based indicators):

- 10. Empirical monitoring data from farm visits of 150 livestock farms could be used to review the suitability of the different animal-based indicators for integration in a conceptually sound multi-criteria assessment of animal welfare.

Solutions suggested to overcome problems in case studies

Climate stability, Italy:

- 1. Using full FADN data can be a cost-effective data source, when it has been collected.

High Nature Value areas, Italy:

- 3. Using proxy indicators could overcome the lack of information, validity of these indicators should be conducted
- 4. The regional level analysis restricts the number of observations. Analysis should be conducted with inter-regional samples when RDP measures are similar across regions.

Diffuse water pollution, Germany (Nmin indicator):

- 8. A less robust matching approach could be applied with limited ability to deal with other factors affecting uptake and selection bias. The current matching of similar farms is still a more robust approach in pairwise comparison compared to results without any matching attempt.

Animal welfare, Germany (animal-based indicators):

- 10. In most cases, where secondary data are not sufficiently available, an ad-hoc approach to sample selection in pair-wise comparisons of means can be applied. More robust comparison group design using statistic-based options to explicitly consider sample selection issues requires the availability and access to animal welfare benchmarking databases (as for example existing in Scotland) or the integration of additional empirical data from farm visits with existing livestock databases at farm level (e.g. the HIT database in Germany). However, those solutions can only be cost-effectively applied in longer-term evaluation contracts.

Synthesis

Qualitative and naïve quantitative approaches were frequently used to construct a counterfactual in the case studies. This is reflected in the problems reported, where lack of data stood out as the most important reason to resort to more naïve methodological treatment of the counterfactual. By no means were the approaches for analyses simple. For example, the carbon footprint approach used in Italy is a complex approach, partly thus contributing to the problem of getting enough data for a full set of comparison groups across different farm types. The choice of indicator thus affects data availability and the possibilities to construct a counterfactual. Essentially this means that the evaluator may need to prioritize the impact indicators available and see the level of counterfactual analysis possible in each case before choosing the method of constructing the counterfactual (unless more than one approach are used). A poor indicator with a good counterfactual may not be preferable to a good indicator with more circumstantial evidence on impact. Implications of different indicator selections and counterfactual applications on the cost-effectiveness of evaluations are explored in Deliverable D7.2 (Wolff et al. (2016)).

Data quality is found often lacking for more elaborate analysis: spatial reference, data on non-participants and, again, administrational data dissemination problems are mentioned as hindrances to better counterfactual analysis. It is notable, that it is not necessarily the lack of monitoring activities that cause problems, but rather the monitoring frequency, resolution, and the handling of current data.

3.1.3 Statistics-based evaluation options – Explicit approach to sample selection

Strengths of the approach

1. Sample selection can be handled
2. Unobserved characteristics can be handled with panel data
3. Evidence-based evaluation with quantitative results
4. Spatially explicit analysis possible

Weaknesses of the approach

1. Gaps in data for non-participants and other comparison groups
2. Gaps in panel data
3. Gaps in links between environmental indicator data and participation data
4. Multiple unobserved intervening factors cause problems
5. Single measure evaluation better handled than programme level evaluation

Problems encountered in case studies

Biodiversity wildlife, Hungary (macro-level assessment):

- 1. Representation of participant-non-participant attributes linked to biodiversity data collection was challenging.
- 2. Due to the size of survey squares micro-macro linkages were difficult to analyse and the assessment of the impact of local environmental circumstances faced difficulties.
- 3. Yearly effects (drought, unfavourable weather conditions during breeding season) may influence the final results of biodiversity assessments.

Biodiversity wildlife, Hungary (micro-level assessment):

- 4. Micro level impact estimates may be biased due to local environmental circumstances.
- 5. Yearly effects (drought, unfavourable weather conditions during breeding season) may influence the final results of biodiversity assessments.

Biodiversity wildlife, Lithuania:

- 6. Wildlife is a large concept.
- 7. Number of non-participants was low in the study area.

Landscape, Greece (land cover change indicator):

- 8. The objective of evaluated measures is maintenance rather than improvement.
- 9. IACS data did not include information for non-participants.

- 10. Method produced quantitative results, but was not able to explain the effects.
- 11. Farm-level data, which is the decision level for participation in the various schemes, was missing, making programme and measure evaluations difficult.

Landscape, Greece (visual amenity indicator):

- 12. Measures of landscape quality were strictly unrelated to the physical features and distinctive characteristics of the landscape.

Diffuse water pollution, Germany (GNB indicator):

- 13. Assumed strengths of having different data sources (e.g. monitoring data, farm accounting data or control data of the fertilizer ordinance) available to create samples (increased scope for sample selection) turned out to be problematic. Comparability and reliability of data sets is limited as nutrient balances are calculated by different kind of stakeholders.
- 14. Structural differences between different data sets exist. Data gaps in particular for non-participants could be confirmed.
- 15. The high variance of single values requires a large sample size. Sample collection is quite expensive.
- 16. While sufficient data were available to apply PSM, only some factors affecting uptake could be considered.

Solutions applied to overcome problems in case studies

Biodiversity wildlife, Hungary (macro-level assessment):

- 1. Spatial analyses and classification of biodiversity survey squares based on the AE uptake contributed to better analyses. This was possible due to the relatively large amount of data available.
- 2. Micro level surveys focused on using baseline data sets of Common Birds Monitoring Program, where the size of the functional unit allows the assessment of micro level impacts.
- 3. Long term data sets allowed carrying out trend analyses.

Biodiversity wildlife, Hungary (micro-level assessment):

- 4. Involvement of naturalness as a side attribute for the analyses helped to filter out the most important external factors and explored external driving forces.
- 5. Long term data sets allowed carrying out trend analyses.

Biodiversity wildlife, Lithuania:

- 6. Choice of specific indicators very closely related to the agri-environmental measures.
- 7. Study area was enlarged to include more non-participants.

Landscape, Greece (land cover change indicator):

- 8. Change is observed in the case of non-intervention. Given that change was easily observed, the construction of comparison groups was possible.
- 9. Non-participant group was constructed using remote sensed data.
- 10. DiD analysis was limited only to the observed changes.
- 11. Macro level evaluation was only based on spatially up-scaled micro results.

Landscape, Greece (visual amenity indicator):

- 12. Evaluated measure explicitly states which vineyards offer high amenity values.
- 12. Common sense interpretation of the observed changes and how these changes may affect the amenity values offered by the traditional vineyards. The categorisation of indicator into three levels was based on arbitrary criteria assigned by the research team.

Diffuse water pollution, Germany (GNB indicator):

- 13. And 14. A smaller sample of one data source was used.
- 15. The cost-effectiveness of different sample sizes has been compared.
- 16. The PSM is still superior to more naïve approaches as it improves the robustness of results.

Synthesis

Several case studies were able to employ a statistical approach attempting to tackle self-selection issues in constructing the counterfactual. The variety of approaches in constructing the counterfactual, some of them based on spatial analysis while others were based on non-spatial data, shows that more elaborate evaluation is possible even with current data sources. However, none of the case studies could employ the PSM-DID method with the available data. This is a symptom of very different data sources across the countries and different cases. Where the tailored regression approaches are effective in using the available data to its fullest, they do not allow easy comparison between case studies. As the statistical methods hinge on data quality, the problems reported are less surprisingly related to data quality. Quality was mentioned as suffering from insufficient data on factors affecting participation to the evaluated measures, linking indicator (and monitoring) data to farm-level actions, lack of data on external factors driving changes in the indicator, and problems in combining multiple data sources[20]. While the data was not perfect, evaluation taking account of sample selection was still possible, albeit difficult.

3.2 General lessons

The case studies following the logic model have found ways to construct a counterfactual, or even sets of counterfactuals with a number of different approaches. The main hindrance for more elaborate evaluation is consistently found in the quality and quantity of data. Surprisingly, the lack of data altogether did not stand out as the only problem. Often the problems were related to the data quality, the spatial and temporal resolution of data did not fit the functional units of evaluation

[20] Case studies using statistical methods reported less often the lack on non-participant comparison group data. This is natural, as statistical methods typically require non-participant data to be usable at all. Thus, when non-participant data was not available, the case studies would have resorted in other evaluation approaches. Generally, however, the lack of comparable non-participant data is a hindrance for evaluators wanting to use statistical comparison methods.

(regional or farm-level) or the temporal scale of the evaluation period. Dishearteningly, there were cases where administration stepped in the way of evaluation, putting unnecessary hindrances on public data, especially FADN use. Such hindrances include providing only aggregated data and physical restrictions on data use (e.g. data tied to location with limited programmes for analyses on-site).

Surprises were found in methodologies that should provide easy counterfactual analyses. A noticeable example was in ensuring the consistency of the results from the Finnish structural model on diffuse water pollution that would have required more time than was available for analysis. On the other hand, the other Finnish case study on climate stability using a continuously updating sectoral model suffered from no problems in analysis.

Generally, the case studies using direct impact indicators (i.e. the actual environmental impact) had more often data availability related problems. Monitoring is expensive, leading to relatively small data repositories which may not cover non-participant farms or areas. Further, actual environmental impacts are often slow processes which may not exhibit during the evaluation period even with good data availability. In these cases modelling approaches, qualitative analysis or using proxy indicators (e.g. indirect pressure indicators) in statistical analysis may be preferable.

The logic model shows promise in finding different approaches for constructing a counterfactual with varying data availability. As some case studies discovered, data quality as initially perceived did not always allow the intended method for constructing a counterfactual, but the logic model could in those cases be used to trace back other methodological options. Additionally the case studies show that sometimes an ad-hoc approach to sample selection may help in improving the evaluation quality from simple comparisons. Ad-hoc approaches rely on the evaluator's ability to understand the underlying selection process to measures. However, such approaches are not easily comparable over time and across measures warranting improved monitoring that specifically aims to help evaluation process.

The structure of the methodological framework for counterfactual development will inform the development of the methodological handbook for the evaluation of environmental impacts of rural development programmes, which will pay particular attention to the practical problems encountered as well as the solutions applied in the public good case studies.

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 (2014) *Report on monitoring and data requirements for counterfactual methods*. Report to the European Commission. Deliverable D3.2, ENVIEVAL project (Project Reference: 312071).

Artell, J, Aakkula, J, Toikkanen, H (2013) *Review of counterfactual methods*. Report to the European Commission. Deliverable D3.1, ENVIEVAL project (Project Reference: 312071).

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).

Pufahl, A, Weiss, X. (2009) (Evaluating the effects of farm programmes: results from propensity score matching European Review of Agricultural Economics 36 (1), 79–101

Wolff, A, Schwarz, G, Osterburg, B, Offermann, F (2016) *Report on the cost-effectiveness of the evaluation approaches.* Report to the European Commission. Deliverable D7.2, ENVIEVAL project (Project Reference: 312071).