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 D5.3

Report on the methodological framework at macro level

Authors: Inge Aalders (JHI), Jane Morrice (JHI), David Miller (JHI), Gerald Schwarz (TI)

- **with input from the AUA, BEF, CREA, LUKE, SZIE and TI teams**

Approved by Work Package Manager of WP5: David Miller (JHI)

Date: December 2015

Approved by Project Coordinator: Gerald Schwarz (TI)

Date: December 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 Acronyms

AE	Agri-Environmental
CF	Carbon Footprint
CGE	Computable General Equilibrium
DiD	Difference-in-Difference
DREMFIA	Dynamic multi-Regional sector Model for Finnish Agriculture
FADN	Farm Accountancy and Data Network
FBI	Farmland Bird Index
GHG	GreenHouse Gas
GNB	Gross Nutrient Balance
HNV	High Nature Value
IACS	Integrated Administration and Control System
LPIS	Land Parcel Identification System
PE	Partial Equilibrium
RD	Rural Development
RDP	Rural Development Programme
USLE	Universal Soil Loss Equation

List of Figures

List of Tables

1 Summary

ENVIEVAL has placed the macro-level evaluation of the impact on the environment of rural development programmes and measures firmly in the context of the necessity to combine micro-level and macro-level evaluations using a robust counterfactual approach to assessing net-impacts. This report describes the macro-level logic model and the results of the case studies, which have tested the evaluation methodologies and the logic model.

Data availability and quality are the over-riding constraints for the development of a robust quantitative macro-level evaluation based on a sound counterfactual. The disparity between data for environmental monitoring and policy uptake creates inconsistencies between micro and macro-level assessments which have an adverse effect on the net-impact assessment. However, it also affects an evaluator's ability to create comparison groups that can support the application of an elaborate statistics-based evaluation.

A robust assessment of comparison groups using elaborate statistics-based methods has been demonstrated to be difficult at a macro level. In most case studies, only a naïve group comparison of with / without and (or) before / after situations could be done relying on a simple mean for the difference-in-difference. Whilst this is mainly caused by data issues, the need to have a large number of different regions to allow for robust econometric assessments of comparison groups at a regional level is also an issue. Macro-level modelling approaches (e.g. applied in the context of the public good soil quality) were able to fill some of the data gaps but relied on established causal links between agricultural practices and the relevant public good. Using elaborate statistics-based methods for the assessment of counterfactuals with comparison groups is more applicable at a micro level.

Two ways of conducting macro-level evaluations of environmental impacts emerge from the experiences of the case-study testing. First, upscaling of micro-level results was possible in most of the macro-level case studies, although statistical representativeness could not always be achieved due to data gaps. However, using GIS-based spatial evaluation approaches potentially provides a solution for consistent aggregation and upscaling of micro-level results to a macro level. Plausibility checks of the results are suggested as a means of reviewing the extent to which the occurrence of indirect effects can be interpreted.

Second, solutions for macro-level evaluations based on specific macro-level modelling approaches were tested successfully. Modelling methods provide different opportunities to fill the gap created by the lack of observational data availability or quality and provide solutions for the evaluation of measures which are implemented area-wide (i.e. not sufficient non-participants exist). However modelling methods often require particular skills and expertise and are not without their data issues. For example, both of the climate change case studies that estimate the impact on GHG emissions are data demanding in their own right. Another approach to deal with data gaps is the use of geostatistical methods to assist in aggregating data, such as the FADN data for the HNV case study in Italy.

Overall, the macro-level logic model has proven to be useful in highlighting key considerations at critical moments during an evaluation, which support managing authorities and evaluators to gradually improve the quality of the evaluation towards a more quantitative assessment.

2 Background

2.1 Rational for the framework

The evaluation of RDP impact on the environment consists of three main components: a sound counterfactual design[1], and assessments at micro[2] and macro levels. These three elements of the framework (shown in simplified form in Figure 1) are linked and, following consistency checks between micro and macro levels, they collectively inform the net impacts of RDP. The framework provides a context and transparency that assists in structuring the assessment by defining sound comparison groups, and includes a check that the results from both micro and macro level are consistent, which can improve the quality of the evaluation.

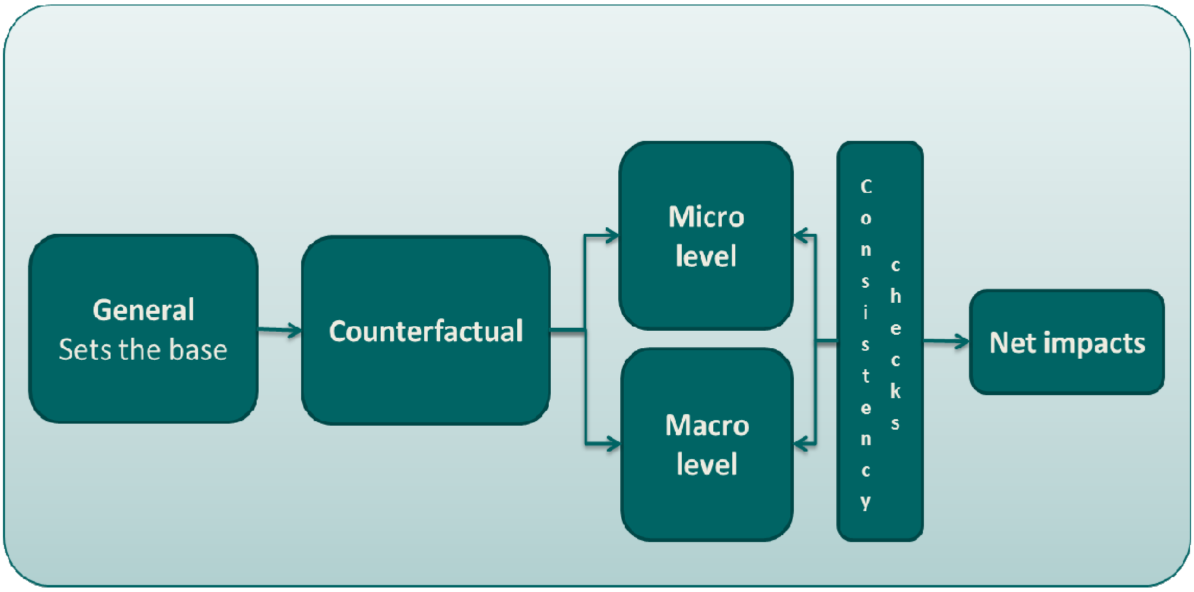

Figure 1: Simplified Evaluation Framework showing the elements of counterfactual, micro and macro levels

Maintenance and changes in the public goods, that RDP objectives aim to achieve, are commonly assessed at a different scale at different levels (see section 2.2). If macro level is defined as beyond farm (i.e. micro) level, most of the public goods are commonly assessed and monitored at macro level. This creates data and methodological challenges for the evaluation. On the other hand, if macro level is Programme level, with micro level being the contract level, then for the assessment of the net impacts it is necessary to understand the causality between changes in public goods and RDP measures as well as the ability of the impact indicator to measure that maintenance or change, either at micro- or macro-level.

2.2 Challenges

2.2.1 Data availability and quality

Data monitoring for both RDP and public good specific purposes take place throughout Europe. Strategies for collecting data are a mixture of EU regulatory and administrative responsibility and national monitoring priorities. Where monitoring data are being collected,

[1] For more detail see Artell et al. (2015) on the methodological framework for counterfactual development.
[2] For more detail see Povellato et al. (2015) on the theoretical and methodological framework for micro-level.

they are not always readily available for analysis or suitable for the preferred methodology[3]. A commonly-cited reason for restricted access is that of confidentiality.

If data are available, the key aspects that determine the quality of the data are:

- Indicators and other variables that cover environmental indicators as well as a wide range of farm / land management variables and wider socio-economic and policy variables.
- Methods with primary data on environmental indicators at farm and field level (note: most commonly used data types are land use and farm management data from IACS/LPIS, FADN, Census and Eurostat databases).
- The sample size / population covered to include an appropriate population in terms of size, coverage and representativeness. Separate consideration of participating farms/areas and non-participating farms / areas.
- Continuity in data formats for non-spatial data (volume and value formats), spatial data (polygon and raster format) and time series of annual data and periodical data for the evaluation period.
- Means for spatial aggregation and disaggregation.
- Consistent integration of multiple data sources.
- Quality and consistency checks of extraction process, data merging, study variables, assumptions.

2.2.2 Scale and levels

The evaluation of RDP impact on public goods relies on our ability to measure change in the public good as a result of RDP implementation. Currently, evaluation relies on the integration of monitoring data that are collected for different purposes, which have different definitions of 'micro' and 'macro level'[4].

This ambiguity about 'macro-level' can be explained by the definition of scale and level (Gibson et al., 2000):

- the term 'scale' refers to *spatial, temporal, quantitative, or analytical dimensions used to measure and study any phenomenon*, and
- 'level' refers to *locations along a scale as units of analysis that are located at different positions*.

Agri-environmental measures, which aim to improve public goods, introduce a number of additional dimensions into the evaluation of RDP impact on the public goods. It means that there are different scales, each with their own set of levels. For example, Figure 2 illustrates two scales common to RDP and two scales for public goods; each of these scales has levels that can be interpreted as micro and macro.

[3] For more detail see Schwarz et al. (2014) on monitoring and data requirements for macro-level methods.
[4] For more detail see Section 2 of Aalders et al. (2013) on the review of macro-level methods and scales.

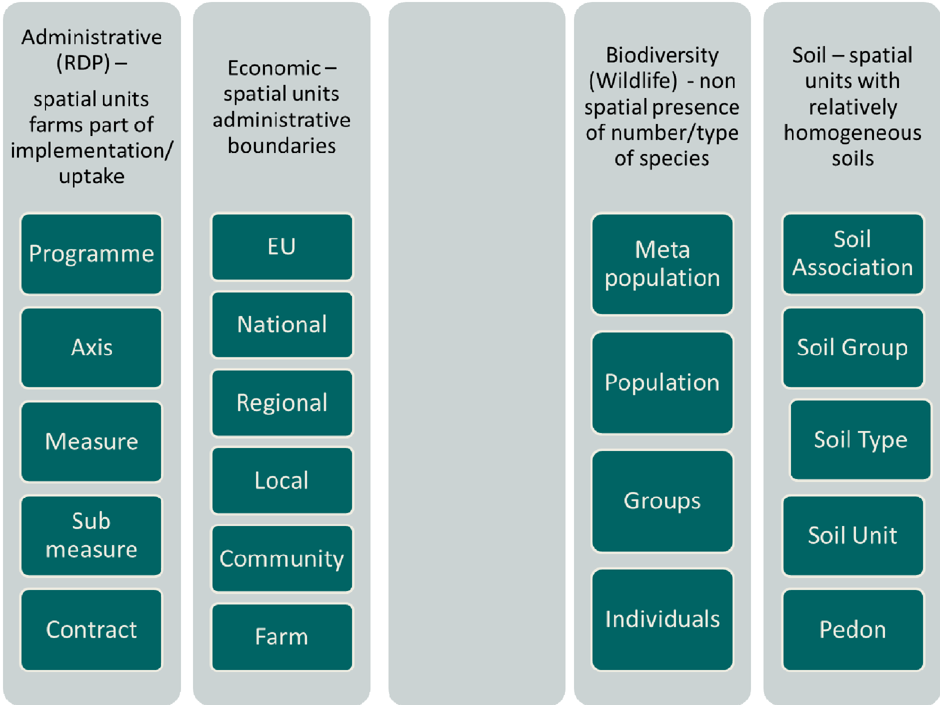

Figure 2: Examples of different types of scales

The micro and macro levels in RDP are linked by the intervention logic which provides a hypothetical trajectory from the Programme, through its objectives, and the measures, to the impact on the environment and thus the beneficiary. These levels are also linked bottom up, through the evaluation logic, from the impact on the beneficiary, created by the measures, to achieve the objectives of the Programme. In other words, the presumed chain of effects links the individual measures (micro level) with the Programme (macro level). In general, scales represent a hierarchy where the within-scale, cross-level interactions will be coherent. This means that there is no confusion about the interpretation of what is meant by the 'macro level'.

With the introduction of environmental components in the RDP, the process of monitoring progress and assessing impact against the set objectives has become considerably more complex. As the example in Figure 2 illustrates, an important factor in the challenges of evaluating the impact of RDP on the environment is that the activities of the RDP incorporate significantly different scales and levels. It means that the measurement of RDP impact may be more appropriate at another level along the scale.

For the Evaluation Framework the relevant macro levels are:

- Regional level – this level is above that of observation of a public good.
- Programme level – this level represents the overall impact of RDP.

The methodologies used in the case studies of ENVIEVAL included examples of both of these levels for the macro assessment. Multi-criteria analysis, particularly agri-environmental footprinting (Purvis et al., 2009; Mauchline et al., 2012), is the most promising approach to assess the complexity of multiple measures. However, there are key challenges to enabling the translation from farm level to regional and national levels. In this context, there are lessons from other methodological approaches for scaling (Ewert et al., 2011) and conceptual integration (van Ittersum and Brouwer, 2010, Ferrier and Drielsma, 2010).

Based on the developments towards integrated assessments, a macro-level assessment should focus on the wider environmental impact which can be defined as regional level and be at different geographic scales and not just on the overall Programme assessment (i.e. effectiveness of investment). This requires the inclusion of spatially-explicit analysis at appropriate levels. Other EU projects (e.g. SPARD[5], FLINT[6], ADVANCED-EVAL[7]) have addressed a number of relevant issues in the effort for integrated analysis, which is facilitated by the use of evaluation levels that are common across scales, e.g. farm is a level shared by administrative, spatial and programmatic scales.

In relation to the overarching evaluation challenges identified and validated at the beginning of the ENVIEVAL project (see for example Deliverable D9.3 and D6.3), particular gaps in the current knowledge are:

- limited evidence of causality between the RDP objectives and indicators
- limited evidence of the ability of indicators to measure impact across and within scales and levels (micro-macro level consistency issues)
- necessity to incorporate the spatial context around participating areas (need to consider the wider local and regional environmental characteristics).

2.2.3 Causality between farm activity and change in public good

Agri-environmental measures aim to support changes in farm management that will have a benefit to public goods. However, it is often difficult to measure the impact that changes in management practices of a single farm have on any of the public goods (in particular climate change, biodiversity, water quality, and landscape).

This is due to:

- The difference in the availability and the collection processes of data for use in monitoring of RDPs and public goods means that there is commonly a mismatch in the quality of data, both spatially and temporally. This makes it difficult to prove direct causality of the RDP action on the public good. This is one of the main obstacles to establishing true causality between measure and public good.
- The CMES indicators are measured along a scale of economic spatial units (farm, regional, national and EU) while each of the public goods have their own scale of measurement (as illustrated by Figure 2), which are used to develop a cost-effective monitoring programme for that public good (e.g. wildlife biodiversity by group, soil by soil type). The difference in monitoring activities poses a challenge to the establishment of causality between farm activity and change in public good, because often the data quality is insufficient for a robust assessment based on a sound counterfactual design and causality.
- Spatial correlation between the land-use management and the public good. This means that the location of the uptake of the measure in relation to the public good is important and therefore that some uptake, due to its location, can have a much higher impact on the quality of a public good than another. This has particular implications for the micro-level indicators. For example, on-farm indicators may not be appropriate for the measurement of the public good or there may be no adequate data

[5] For more information on SPARD : http://project2.zalf.de/spard/
[6] For more information on FLINT : http://www3.lei.wur.nl/flint/index.html
[7] For more information on ADVANCED-EVAL : http://cordis.europa.eu/result/rcn/47026_en.html

available to establish direct causality between measure and public good. In this context, the introduction of alternative indicators that measure any spatial relationship between the uptake of measures and the public good may provide a better indication of impacts on public goods such as water quality, biodiversity and landscape than aggregating existing indicators along the RDP scale.

2.2.4 Micro – macro level consistency

A problem arises with consistency when the demand for an evaluation at one level does not correspond to the required spatial or temporal scales, or the requirements of the programme. This is because, while the intervention is at the farm level, the evaluation is, in most cases, at the ‘contract’ level.

A simple aggregation of the effects from micro level to macro level is not always adequate or correct to enable a description of effects at the macro level. In the context of public goods, the necessary observations (e.g. temporal and spatial sampling strategies) do not support an assessment of impact at micro level.

There are five methodological approaches considered by Lukesch and Schuh (2010) for dealing with the micro – macro consistency at regional or programme levels which are:

1) *Statistical/econometric method for*:
   a) Comparison of programme participants (e.g. farms, food processing enterprises, specific rural communities, etc.) with equivalent non-participants;
   b) Comparison of programme areas with comparable non-programme areas (or with other areas characterised by a different intensity of the programme in question); via estimation of direct and (under certain conditions) specific indirect programme effects

2) *Regional input-output models;*

3) *Micro-macro models (including Computable General Equilibrium [CGE] framework);*

   The descriptions of micro-macro models contain the first references to consistency checks, but only for economic variables. However, in the rest of the text, there is no such reference; instead the difficulties of scaling (both directions) are noted.

4) *System dynamics modelling (especially for assessing impacts in the environmental field);*

5) *GIS based assessment tools*

In the context of public goods and the delivery of environmental goods by RDP, of the five approaches presented GIS-based tools in combination with economic modelling are the most appropriate for micro-macro consistency:

> *“Modelling approaches are especially in the field of environmental impacts limited in their explicatory power. RD measure effects on the environment are more likely to be seen in the real world rather than an abstract model. Thus the observation of changes in territories, where RD measures have been applied are a common practice to assess the impact on the micro scale and then add these territorial pictures up to the macro scale as well.”* (Lukesch and Schuh, 2010).

The key problem regarding micro-macro consistency is in relation to data. Indicators most directly related to environmental outcomes are collected at the regional or national level. While this data is important in measuring the end results, it normally does not link farm practices with the efforts made at all levels that enable the development of more sustainable agricultural practices.

High-level aggregated indicators and proxy indicators can ‘hide’ impacts and usually do not give an accurate picture of the farming or environmental conditions at a specific location. However, robust impact models which link farm level (practice) with landscape level (impact) can be designed to avoid inconsistencies. Triangulation, i.e. the use of more than one method for the same assessment, is another method of increasing consistency.

Possible solutions to deal with issues associated with data include ‘logical’ or conceptual approaches and not technical ones. These are:

- a common framework or common elements of a framework;
- use of comparable indicators and metadata;
- collaborative capacity-building and data platforms that allow data submission by a distributed network of actors, potentially including farmers;
- clarity on impact pathways across scales and actors;
- attribution of impact to various drivers or pressures at the different levels;
- complementary capacities;
- platforms and openness for cross-scale dialogue;
- comparable legislative requirements;
- a clear incentive mechanism and cost-sharing across the many levels; and
- multi-scale overview, audit pressure and mechanisms.

Finally, a possible solution for consistency checks of the net impact assessment is the comparison of aggregated micro-level results with the results of specifically designed macro methods

3 Macro-level logic model.

3.1 Spatial context of the assessment

For the impact of RDP measures on public goods, such as biodiversity, water quality, soil, and landscape, the geographic location of actions can be highly significant. In relation to public goods such as climate change and animal welfare, the activities are less, or not, spatially correlated with impacts. In the case of biodiversity, the effective positive impact of RDP on the population of declining wildlife species and available habitat relies on the spatial relationship between existing biodiversity and the implementation of specific RDP measures. Similarly, the relationship between agricultural activities and the quality of soil and water are strongly spatially related which means that the same measure can have different levels of impact depending on where it is implemented within a catchment.

3.2 The three logic models for macro level

The methodological framework has three different macro-level logic models, one for each of the three types of counterfactual[8] designs. The steps within the three macro-level logic models are the same:

Step 1: Definition of unit of analysis and the consistency of the indicators

a. Unit of analysis [9]
b. Consistent indicators

Step 2: Creation of consistent spatial data

a. Spatially explicit data
b. Appropriate resolution

Step 3: Selection of Counterfactual Approach

a. Evaluation Option without Comparison Groups
b. Qualitative and Naïve Quantitative Evaluation Options – ad hoc approach to sample selection
c. Statistics-based Evaluation Options – explicit approach to sample selection

Step 4: Net Impact Assessment

a. Micro-macro consistency
b. Net impact assessment

3.2.1 Step 1: Definition of unit of analysis and the consistency of the indicators

Following the selected counterfactual approach, with the available data, the first step in the macro-level logic model is to identify a unit of analysis (**Fehler! Verweisquelle konnte nicht gefunden werden.**). At this stage the evaluator needs to consider the most appropriate

[8] See for more detail D3.3 Report on the methodological framework for counterfactual development
[9] **Unit of analysis** is the most elementary part of the phenomenon to be analysed and its definition influences the design of analysis and data collection (Frankfort-Nachmias and Nachmias, 1996). For more detail see Section 2.4 of D4.3 Report on the theoretical and methodological framework for micro-level

unit of analysis given the available data and the identified counterfactual approach. In addition, the evaluator examines, as part of the micro-macro consistency, whether micro-level results can be integrated into the macro-level assessment (i.e. scaled up).

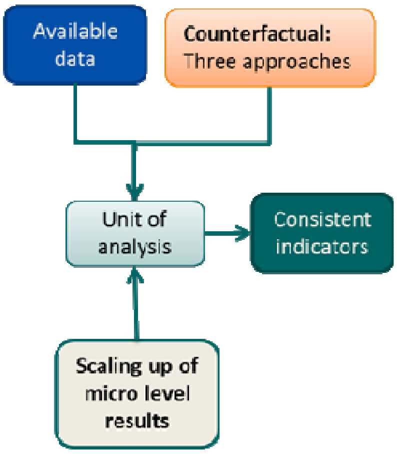

Figure 3: Step 1 Definition of unit of analysis and the consistency of the indicators

For the selected unit of analysis, the evaluator considers whether the selected indicators are consistent. *Consistent indicators* are those for which micro-macro consistency can be demonstrated (see section 2.2.4).

3.2.2 Step 2: Creation of consistent spatial data

The availability of spatially-explicit data is important for the assessment as, potentially, it can assist in the provision of quantitative evidence for causal relationships between RPD and impact on public goods. The availability of spatial data determines the type of analysis that can be conducted for each of the three counterfactual approaches.

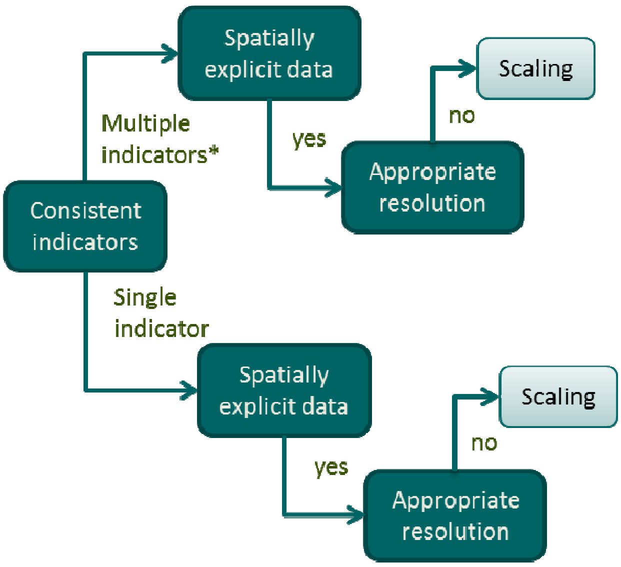

Figure 4: Step 2 Creation of consistent spatial data

At a macro level, it is desirable that the impact assessment is able to incorporate multiple indicators as well as single indicators. There are methods that give the evaluator the option to consider multiple indicators for an impact assessment. However, these differ from the method for single indicators. Therefore, with the consistent indicators, the evaluator has to decide whether the assessment will be based on a single indicator or multiple indicators. For the application of multiple indicators, the evaluator will use the macro-level logic model for a single indicator assessment, separately with each indicator, as well as for the multiple indicator assessment. The use of multiple indicator methods for the impact assessment is considered to lead to a more robust net impact assessment.

Ideally, an assessment will take place with data that have the same *spatial resolution.* Where there are spatial data at different spatial resolutions (e.g. spatial support for economic actors at a municipal scale, water quality data ta a river basin level, soil data for individual soil units), an extra step is required prior to analysis, in which the spatial resolution is harmonised either through up-scaling or down-scaling methods to a single resolution.

3.2.3 Selection of Counterfactual Approach

The three different counterfactual approaches lead to a selection of different evaluation methods, which have specific capabilities to assess the impact of RDP on a public good given the unique combination of indicator and data availability conditions with which the evaluator is working.

The methods for ***evaluation options without comparison groups*** (Figure 5) can be divided into those using spatial and non-spatial data for the modelling process. For non-spatial data of single or multiple indicators, *Computable General Equilibrium (CGE) and Partial Equilibrium (PE) modelling frameworks* can be used to deal with situations without comparison groups. Temporal dimensions of environmental impacts are directly incorporated in the dynamic modelling framework and policy impacts are quantified based on before and after simulations. The modelling framework provides the flexibility to simulate different counterfactual scenarios and the regional differentiation enables the interpretation of indirect effects at a macro level, such as displacement and substitution effects. Care must be taken with respect to the assumptions applied to implementation of the policy measures in the modelling framework to ensure that the causal relationships of the policy measures and related land-management changes are theoretically sound.

For spatially-explicit data, there are three different modelling methods that can support single and multiple indicators. These methods each have unique contributions to make to the evaluation. For example, *spatial econometrics* is specifically able to disentangle the external impacts of other intervening factors from the environmental changes which can be directly attributed to the policy measure or programme. However, spatial econometrics has not been tested in the ENVIEVAL project, as a separate FP7 EU project focussed solely on the application of this method in RDP evaluations (SPARD project).

Reinhard and Linderhof (2013) conclude that the spatial econometric models tested provide a suitable method to assess environmental impacts of RDPs at a macro level. This allows the incorporation of counterfactuals through analysing regions with different spending on the measures and different historic and prospective pathways of development of biodiversity and water quality. However, substantial data requirements for these methods can limit an application for macro-level evaluations of environmental impacts.

Finally, *landscape metrics* explore the causal linkages and the consistency of micro-macro linkages. The use of changes in landscape spatial metrics of land cover and use associated with RDP measures compliments that of changes in the visibility of land cover and use associated with RDPs, for interpretation with respect to landscape character. The combination enables the assessment of a broader set of net effects and better captures the complexity of environmental relationships with respect to the character of the landscape and thus the public good. Local environmental characteristics are included, and explicit analysis of micro and macro levels are consistently combined.

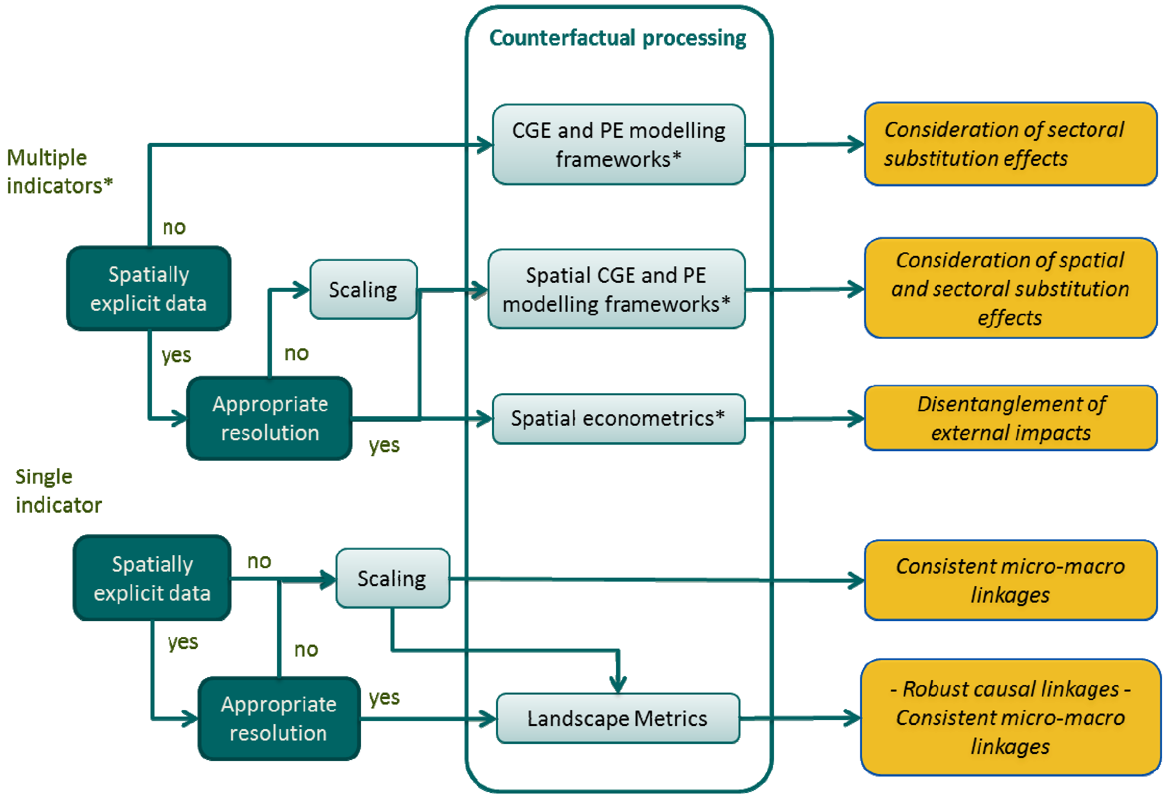

Figure 5: Step 3a Evaluation option without comparison groups

There is a range of different methods for ***Naïve Quantitative Evaluation Options*** with an ad hoc approach to sample selection (Figure 6). In the case of multiple indicators with no spatial data, methods such as *ecological foot-printing and multi-criteria evaluation* allow the evaluator to examine non-spatial and spatial macro-level heterogeneity. *Multi-functional zoning* is a spatial multi-criteria analysis which can consider the spatially-correlated heterogeneity of multiple indicators.

In the case of a single indicator without spatially-explicit data, *hierarchical sampling* is the method use to explore the consistency of micro and macro linkages. For spatially-explicit data, *spatial statistics* can assist in establishing robust causal linkages and examine consistent micro-macro linkages. For specific public goods such as landscape and HNV, there are specific *spatial statistics (e.g. landscape metrics)* that are established methods of assessing these public goods and are being examined for use in the RDP evaluation process.

The most comprehensive assessment is ***Statistics-based Evaluation Options with explicit approach to sample selection*** (Figure 7). The methods used in this counterfactual approach are able to use single or multiple indicators for the assessment. With access to good quality data, the evaluator will be able to build the impact assessment on robust causal links, constituent micro and macro linkages and disentangle RDP impacts from external impact.

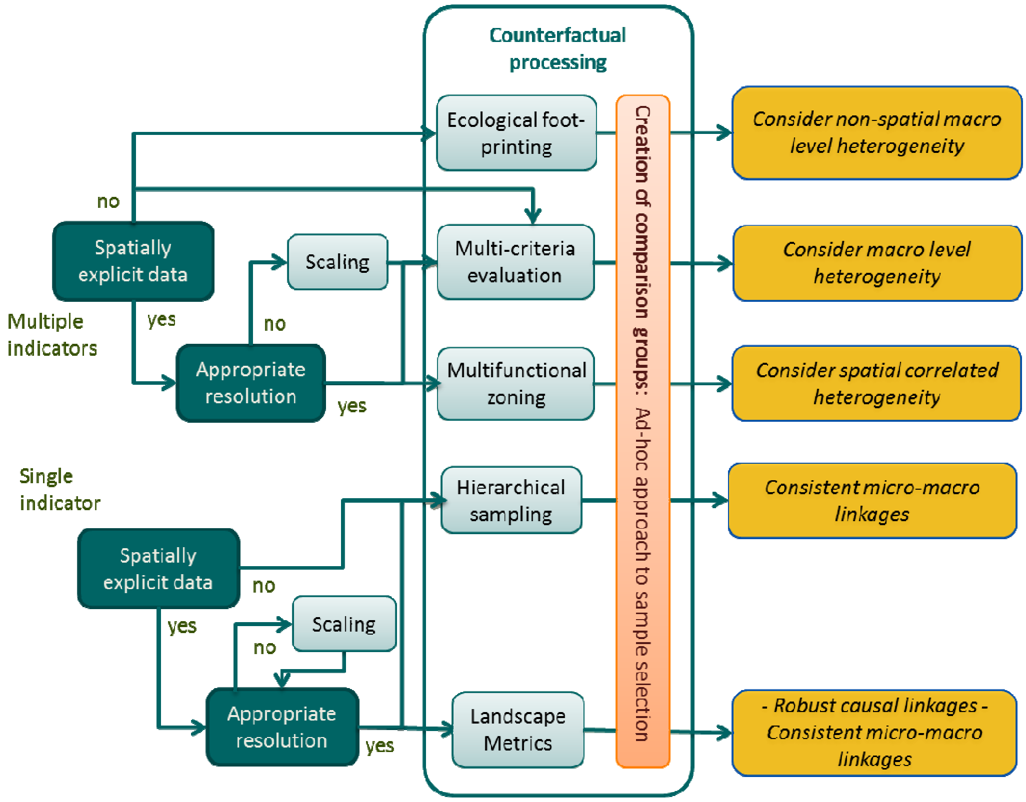

Figure 6: Step 3b Naïve Estimates of Counterfactual

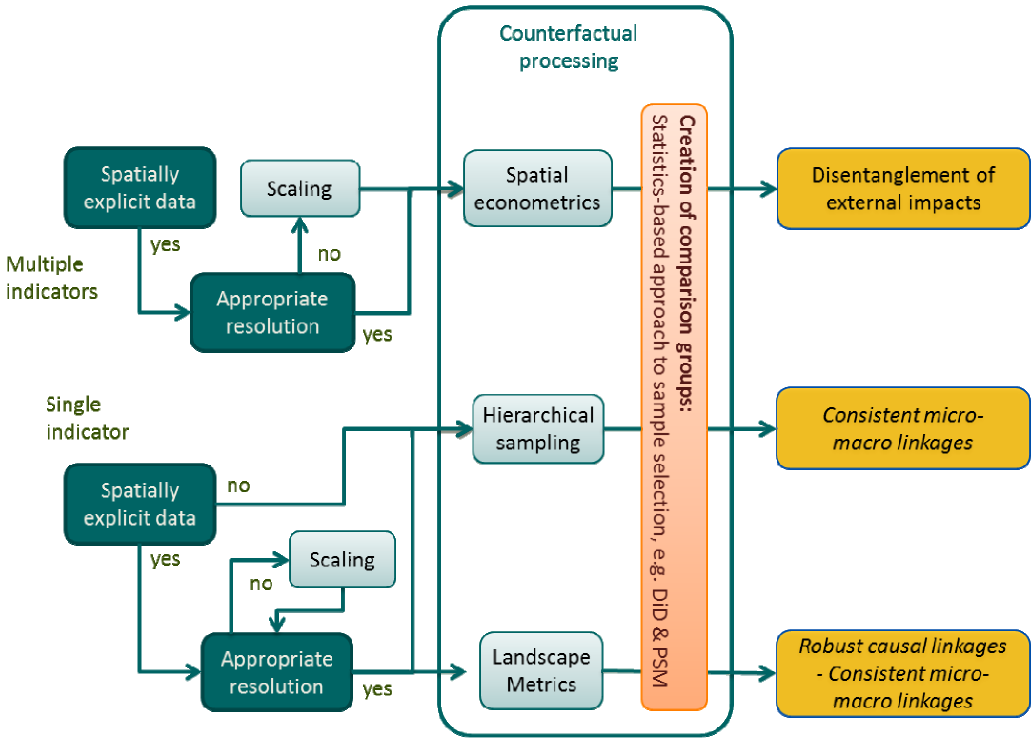

Figure 7: Step 3c Elaborate Statistics-based Evaluation Options

3.3 Contributions to net-impact assessment

The overarching goal of the evaluation is to assess net-impacts of the RDPs. The EU evaluation concept strictly differentiates between programme outputs (physical units), results of measures and combinations of measures under the different focus areas, and programme impacts at macro level on participants and non-participants. While results reflect net-effects on participants or participating areas, which are supposed to be measured with result indicators, programme impacts include direct and indirect effects on participants and non-participants. Indirect effects at macro level include substitution, displacement and multiplier effects. Basically, programme impacts reflect medium/long-term effects beyond the immediate effects on participants (or participating areas) that can be observed at community, regional level or programme area.

The EU concept of taking into account the different direct and indirect effects as part of the net-impacts is complex and requires a triangulation of different evaluation methods across micro and macro level. However, given the different methodological challenges and data issues faced when making environmental evaluations, the main objectives of the methodological framework is to identify specific contributions of methods and approaches moving closer to a theoretically sound net-impact assessment.

Examples of particular contributions of the selected evaluation methods and approaches at a macro level which address the main challenges are highlighted in the yellow boxes in Figures 6 to 8, and are explained in Section 3.2.3. In the context of macro-level indirect effects, one of the main contributions of using dynamic general or partial equilibrium models, for example in the assessment of climate stability impacts of RDPs, is the explicit consideration of indirect effects such as substitution effects occurring within and between different land-use sectors as well as capturing displacement effects of GHG emissions between different regions. However, the explicit consideration of the indirect effects in macro-level assessments requires spatial and non-spatial data of sufficient quality and quantity to enable the application of specific macro-level modelling approaches or spatial econometric approaches.

Another key contribution to a net-impact assessment is the improvement of the consistency of the results across micro and macro levels. An example in Figure 7 is the use of landscape metrics, e.g. to assess landscape or HNV impacts of RDP measures. Landscape metrics includes explicit analysis of micro and macro levels and improves the consistency of the micro-macro linkages of the net-impact assessment.

Macro-level evaluations can build on the upscaling of micro-level results or apply a separate macro-level evaluation approach. The latter would include the combination of a ‘top-down’ macro-level approach (e.g. models with national or regional coverage) evaluating programme impacts with a ‘bottom-up’ micro-level approach assessing net-effects of different measures or measure combinations. Both cases require plausibility and consistency checks to be carried out in relation to the data used as input and the comparison of the results at different levels. Section 3.4 explores the plausibility and consistency checks in more detail.

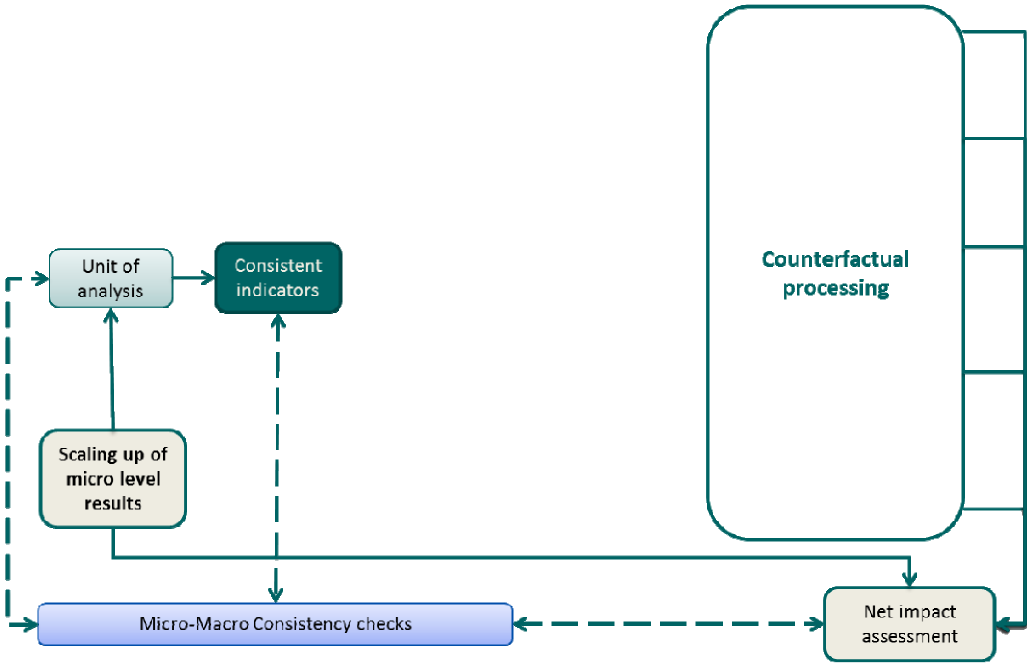

Figure 8: Step 4 Net impact assessment and micro-macro consistency

3.4 Consistency with micro level

The evaluation framework includes a number of points in the process where the evaluator should examine the consistency between the micro and macro levels. These micro-macro consistency checks ensure that the final results are robust and consistent.

There are two main points in the evaluation framework that require a micro-macro consistency check:

1. The selection of the unit of analysis and the identification of the consistency of the selected indicators (step 1);
2. The assessment of the net impact (step 4).

When considering the unit of analysis for a macro-level assessment, the evaluator will take into account the micro-level assessment and its results. The unit of analysis should reflect a scale of the assessment which can be applied at micro and macro levels. Simple aggregation of farm-level (micro level) results to larger areas, possibly along a different scale (i.e. catchments, landscape types, as well as administrative), may not have consistent boundaries, due to farm boundaries not falling within the unit of analysis for macro level. For example, farms may be in more than one catchment/landscape type, while administrative boundaries may not be consistent with the public good boundaries (i.e. catchment boundaries).

For the net impact assessment, it is important that the results of both micro and macro-level assessment are consistent. They are accepted as consistent if results of these assessments show the same trend in relation to impact, even if the evaluator has used different indicators or methods for the assessments.

The main challenge for the consistency with micro level is the causality between farm-level action (micro) and beyond farm boundary change. By incorporating the consistency checks in the evaluation framework both at the beginning and the end of the process, evaluators are reminded at critical moments to integrate the micro and macro-level assessment that will benefit the quality of the net impact assessment. Environmental impacts consist of direct and indirect effects and are driven by different intervening factors at micro and macro levels.

Reviewing the intervention logic and causal relationships between the measures, the required changes in land management practices and environmental change needs to achieve clarity on impact pathways across scales, levels and actors. This is an important conceptual basis for the assessment of the plausibility of the results at micro and macro level. In many cases, in particular due to data gaps, this can only be done in a qualitative way by reviewing the importance and expected direction of change that different indirect effects or other intervening factors might have at micro and macro levels. The key question is whether or not the conceptual basis or framework can be used to explain the differences between micro and macro-level results (i.e. an attribution of the differences in the results or impacts to various drivers or effects at the different levels) and thus in the validation of their plausibility.

In addition, the triangulation of different methods and approaches used to evaluate impacts at micro and macro levels can be used to validate the consistency of micro and macro-level results. Results of macro-level impacts based on the upscaling of micro-level results can be compared with macro-level impacts based on the application of a specific macro level method or approach (e.g. a macro-level modelling approach or a specific calculation of indicators at macro level). For example, up-scaled micro level results of water quality impacts of RD measures at farm level (using the indicator GNB) can be compared to results of an assessment of GNB at catchment level. However, the combination of a bottom-up approach based on micro-level evaluations of the different RD measures at micro level (followed by an upscaling of results) and a top-down approach of macro level using a specific macro-level method to assess RDP impacts requires sufficient resources and longer-term evaluation contracts. In other cases, a triangulation of methods will probably rely on qualitative approaches to validate quantitative evaluation results such as focus groups or stakeholder / expert interviews.

4 Results from the case studies

4.1 Synthesis of the logic model application

In this section the experience gained with the use of the logic models in the case studies is captured. Not all the case studies conducted a macro-level assessment for the public good being tested. Three case-study approaches are only tested at micro level (Table 1): Corncrake density (BW-LT), GNB+water use/ha (WQ-DE) and Integrated animal-based indicators (AW-DE).

Most of the approaches tested in the case studies use spatial data, with only three unique approaches that do not use spatial data. The macro-level functional units/unit of analysis varies significantly between the different approaches from country to public good. The majority of case studies carrying out a macro-level assessment tested upscaling approaches of micro-level results.

Table 1 Summary of the macro-level 'paths' through the logic model – leading to a number of different macro-methods

Case study	Counterfactual	Upscaling Micro-level results	Functional unit	Spatially explicit data	Selected approaches
BW-HU	Naïve counterfactual	y	Country	Y	Farmland Bird Index
BW-LT					Corncrake density/ multiple regression (micro level only)
CC-FI	Without comparison groups	n	Country	N	GHG emission and general equilibrium model
CC-IT	Naïve counterfactual	y	Region	N	Footprint approach
HNV-IT	Naïve counterfactual	y	Region	y	HNV score and multi-criteria assessment
HNV-LT	Naïve counterfactual	y	Region	y	Ecotone diversity and spatial statistics
L-GR	Naïve counterfactual	y	Landscape	Y	Land cover change/visual amenity and spatial analysis
L-SC	Naïve counterfactual	n	Landscape/ region	Y	Landscape structural indicators and spatial statistics
L-SC	Naïve counterfactual	n	Landscape/ region	Y	Visibility of change in land use and cover
L-SC	Naïve counterfactual	n	UAA/ region	Y	Natura2000 and spatial analysis
SQ-HU	Naïve counterfactual	y	Country	y	Soil organic matter content and sampling method
SQ-SC	Naïve counterfactual	n	Sub-catchments	y	Soil Carbon (in tonnes/ha + mean in arable land in megatonnes) and Geostats Modelling (InVEST)
SQ-SC	Naïve counterfactual	n	Sub-catchments	y	Soil Erosion and Geostats Modelling – USLE (InVEST)
WQ-DE					GNB+water use/ha and biophysical modelling (micro level only)
WQ-DE	Naïve counterfactual	y	Water protection area	n	Mineral N content and pairwise comparison and regression analysis
WQ-FI	Without comparison groups	n	Representative farm	y	Nitrogen reduction and biophysical structure modelling
WQ-GR	Naïve counterfactual	y	Specific site of the NVZ of case study area	y	GNB+water use/ha and biophysical modelling
AW-DE					Integrated animal-based indicators and multi-criteria (micro level only)

Although efforts were made to include samples of counterfactual designs based on elaborate statistics in the case studies, this proved not feasible. The approaches for most case studies were limited to a naïve counterfactual design with an ad-hoc approach to sample selection. In addition advanced modelling approaches without comparison groups were used. The following sections provide some background to why case studies were unable to include more advanced counterfactual designs.

4.2 Synthesis of the experiences (positive and negative)

The following is a summary of the experience for each of the public goods, which was recorded by indicator and method tested[10]. The summaries focus on the issues that have arisen during the case studies in general, and more specifically in relation to the data requirements and the methodology.

Biodiversity Wildlife Hungary - BW-HU

<u>Positive experience</u>

- There is good quality data for a relatively long timescale.
- The sample data available (bird census) for the selected macro-level indicator (Farmland Bird Index - FBI) are representative for use at a country level and provide the possibility of combined before-after and with-without comparisons.
- The macro-level data available from 1999 to 2014 support a robust long-term biodiversity assessment.
- FBI has proven to be a robust indicator of farmland biodiversity.

<u>Negative experience and limitations</u>

- Tests rarely involve the assessment of additional intervening factors (environmental, farmer behaviour), which may cause interpretation challenges.
- The formation of participant and non-participant groups faced challenges, as spatial selection of biodiversity survey points have not followed the spatial distribution of parcels contracted under AEMs or other rural development measures.

Biodiversity Wildlife Lithuania - BW-LT

The Lithuanian case study has not conducted a macro-level assessment.

Climate Change/Climate Stability Finland - CC-FI

This case study is a macro level only assessment based on the DREMFIA sectoral model, CO_2 equivalent GHG-emission measures.

<u>Positive experience</u>

- The DREMFIA sectoral economic model allows a number of before-after scenarios to be simulated. That flexibility allows the evaluator to decide which scenarios of a before-after comparison are required for the counterfactual assessment.
- The DREMFIA sectoral economic model uses a wide variety of data (input and output prices, demand for agricultural produce among others) that are continuously collected and used to updated into the model.
- The model provides a methodological solution for situations where non-participants do not exist (or only in very small numbers).
- The model allows interpretation of displacement and substitution effects.

[10] For more detail see annex of D4.3 Report on the theoretical and methodological framework for micro-level

Negative experience and limitations

- Data are continuously collected and updated in the model, which means that unexpected changes in data availability may be problematic for model use.
- The model provides a strictly macro-level assessment that builds on an aggregate micro-level farm response.
- This macro-level model does not identify local impacts.
- Training is needed to use the model.

Climate Change Italy - CC-IT

Carbon Footprint (CF) used for this case study is a well-established method to estimate carbon emissions from functional units having different structural and management characteristics. It is conducted at process or farm level (i.e. micro-level approaches) which is aggregated to a macro level.

Positive experience

- The aggregation of micro-level results to a regional level has been realised using coefficients to include all the crops not analysed at micro level.

Negative experience and limitations

- Statistical representativeness has not been verified.

High Nature Value Italy - HNV-IT

Positive experience

- The HNV indicator requires specific data on land cover, intensity of the farming systems and presence of wild species linked to farmland.
- The aggregation of FADN data at a regional level has been realised using geostatistical interpolation (Kriging method) to define the probability maps on the regional distribution of HNV from farm-level data.

Negative experience

- The data available are not sufficiently exhaustive, either in terms of the range of species covered, geographical coverage and ecological diversity, and they are not updated with sufficient regularity.
- RDP participants in FADN samples at regional level are not sufficiently representative for aggregation to macro level and use of elaborate statistics-base models.
- Macro-level analysis has only been carried out for one year.

High Nature Value Lithuania - HNV-LT

Positive experience

- The method and spatial IACS data can be used to establish robust causal relationship at a macro level.
- Good resolution spatial data on micro level do allow upscaling to a macro level.

Negative experience

- The data available only allow a quantitative assessment of the RDP impact on the extent of HNV but not on the (ecological) quality of HNV.

Landscape Greece - L-GR

Positive experience

- Macro-level assessment is based on the aggregation of indicators calculated at a land parcel level through the reclassification of land-cover polygons.
- The data can be processed with standard GIS software
- Macro-level results can be achieved through up-scaling of spatially explicit micro-level results.

Negative experience

- The assessment was limited to observing change.
- Indicators include quantitative information but are not able to explain the effects.

Landscape Scotland - L-SC

Positive experience

- The methods tested have micro-macro linkages using theoretically based indicators which can be applied to robust spatial units which can be interpreted with respect to their contribution to landscape character.
- Land-cover monitoring data provide a detailed basis for the assessment of landscape structure indicators (baseline assessment). Land-use data recorded as part of IACS data base are a valuable source of information.
- The methods tested can be used to generate individual indicators, which can be used in combination to assess their suitability for use in RDP net-impact assessment.

Negative experience

- Land-cover monitoring is infrequent and out of sync with RDP programme.
- IACS land-use data have gaps in relation to non-agricultural land.
- Reporting of IACS land-use data is not as accurate as land-cover monitoring in relation to non-agricultural land use which has an adverse n impact on the landscape structure. Land cover data (CORINE) may be more suitable.

Soil Quality Hungary - SQ-HU

Positive experience

- The sampling approach is based on a large number of soil samples.
- Modelling approaches can be used effectively to utilise the data available for assessment

Negative experience

- Time for evaluation was short. Farms were not selected according to factors which could influence impact at a macro level. There is no comparison of participating and non-participating farms where all the influencing factors are restricted to a minimum. Selection bias in the comparison of farms could only be addressed to a limited extent.

- USLE model has a temporal resolution of one year.
- CLUE model works at NUTS 2 level that is mainly suitable for a Europe-wide comparison.

Soil Quality Scotland - SQ-SC

This case study is conducted in a data poor environment. National-level soil data monitoring is too coarse for RDP impact assessment. Therefore a macro-level modelling approach has been used for the impact assessment.

<u>Positive experience</u>

- The modelling approach is based on a robust and well-documented theoretical model (USLE) and relationship between soil carbon and land use.
- The results of the models are summarised and represented at a sub-catchment level, due to the uncertainty in the underlying data and analysis.

<u>Negative experience</u>

- Currently, soil monitoring takes place only at a national level and not in a relevant temporal dimension for RDP assessment.
- The data are unable to explain participation/non-participation.

Water Quality Germany - WQ-DE

<u>Positive experience</u>

- Monitoring data are available for use in aggregation within water protection areas (Nmin).
- The availability of monitoring data at different levels of detail for two time periods allows an analysis of micro-macro consistency (Nmin).
- External assumptions have been implemented to improve consistency between results at micro and macro level (GNB).

<u>Negative experience</u>

- Estimations for the macro level could be undertaken but a detailed location analysis is not possible due to limited data availability (GNB).
- Aggregation of data and limited information on farm structural data inhibits sound statistical analysis (Nmin).
- Statistical representativeness between micro and macro-level results has not been verified. Only estimations of environmental impacts at macro level were possible (Nmin).

Water Quality Finland - WQ-FI

<u>Positive experience</u>

- The modelling approach allows assessments based on different counterfactual designs.

- The model uses transfer functions that best describe the environmental condition of the area.
- The structural model uses micro-level data to construct a representative farm at macro level.
- The modelling approach requires quantitative and causal link for the indicator to the unit of analysis, for which existing large databases and previous research supports interpretation of causal links.

Negative experience

- The modelling and the transfer functions used are not spatially explicit.
- The modelling approach is not easily applied by the non-expert, with the model requiring updating consistent with the requirements and restrictions of the AEP. Thus updating results with new data can be time intensive.

Water Quality Greece - WQ-GR

Positive experience

- Aggregation of the results at a micro level based on (land parcel data) could be undertaken at a regional level of the whole case study area due to the availability of spatially explicit land parcel data for the whole area.

Negative experience

- Due to the lack of IACS data at different time points no DiD approach could be applied. However, the availability of annual IACS data should be possible for evaluators allowing a DiD approach to be used.
- Farm level, which is the decision level for participation in the various schemes, was missing in the establishment of micro-macro linkages.
- The micro macro consistency check revealed a serious issue since the differences of the outcomes when using the scaling up of micro results and the macro approaches have not been insignificant. On the contrary for some soil classes they seemed to be quite important, resulting, in cases, to a reverse direction of the difference between participants and non participants.

4.3 Synthesis of general lessons and emerging data issues

Generally the availability and quality of data are the most important factors in the evaluation. They determine the quality of the macro-level assessment and the counterfactual design. Indeed data affect all steps of the macro-level logic model and they are central to all the identified challenges for macro-level assessment.

The results of the case studies reflect the mismatch between the data required and the data available to assess the impact of RDP on a range of public goods. The mismatch is largely due to differences in scale between public good and the RDP.

To resolve the challenges posed by data quality and availability, it is important to consider the complementarity of monitoring objectives between the different levels and to use that complementarity to consider modifications to the monitoring strategy of both RDP and the public goods in a way that it benefits both levels. However, field observational data as used in case studies for biodiversity and soil illustrate that the costs of generating sample data suitable for the creations of robust counterfactual are challenging.

The results provide suggestions for improving the current situation beyond obtaining more data, which is not necessarily the answer. Data collection is expensive, and monitoring aims to have the minimum amount of data necessary to provide the best quality assessment for an intended level. Effective monitoring systems have clear objectives that define sampling strategies and reporting scales to achieve acceptable levels of uncertainty in the results.

A robust assessment of comparison groups using elaborate statistics-based methods has been proven to be difficult at a macro level. In most case studies, only a naïve group comparison could be done. This is mainly caused by data issues, with the need to have a large number of different regions to allow for robust econometric assessments of comparison groups at regional level also being an issue. In the SPARD project, spatial econometrics were successfully tested at macro level using regions across several member states and programme areas. Normally, however, evaluation contracts require specific assessments of the programme area and data access for a single contractor / evaluator across several member states or programme areas is likely to become even more difficult. The use of elaborate statistics-based methods for the assessment of counterfactuals with comparison groups is more applicable at micro level.

Two ways of conducting macro-level evaluations of environmental impacts emerge from the experiences of the case-study testing. First, upscaling of micro-level results was possible in most of the macro-level case studies, although statistical representativeness could not always be achieved due to data gaps. But using GIS-based spatial evaluation approaches potentially provides a solution for consistent aggregation and upscaling of micro-level results to macro level. Plausibility checks of the results are suggested to review to what extent the occurrence of indirect effects can be interpreted.

Second, solutions for macro-level evaluations based on specific macro-level modelling approaches could be successfully tested. Modelling methods provide different opportunities to fill the gap created by the lack of observational data availability or quality; however modelling methods are not without their data issues. For example, both of the climate change case studies that estimate the impact on GHG emissions are data demanding in their own right. Another approach to deal with data gaps is the use of geostatistical methods to assist in aggregating data, such as the FADN data for the HNV case study in Italy.

The case studies for soil and water quality used modelling approaches to overcome the data limitations. Some of these models used for soil quality are more commonly used for the prediction of the implications of climate change rather than measuring the impact for RDP in an ex-post assessment. However, their close causal link with land-use change creates an opportunity to use actual land use for an application in an ex-post assessment. However, models have their own limitations which can range from the effort for their parametrisation to validation of the modelled impacts with actual impacts.

The ENVIEVAL case studies have largely used established causal links in the modelling approaches. These causal links are usually within land use and the public good rather than links between RDP action and the public good. Hence the causal links are embedded within the modelling approaches rather than actually being measured or established based on monitoring data.

In the context of assessing landscape or HNV impacts, landscape metrics include explicit analysis of causal relationships of changes in land cover and use associated with RDP measures using a set of coherent indicators at micro and macro level. This improves the consistency of micro-macro linkages of the net-impact assessment. In general, establishing a good causal link requires statistically-robust data for different possible combinations of public good and RDP activities. Existing data such as FADN, IACS land use and CORINE, in addition to targeted environmental monitoring, are valuable sources of information but the challenges of integrating these data sources commonly prohibits the creation of effective comparison groups for the RDP assessment.

4.4 Consistency micro –macro level

As outlined in the Introduction, the macro-level methodological framework pays particular attention to addressing improved consistency of micro-macro linkages as one of the main evaluation challenges for the assessment of environmental net-effects of RDPs and their different measures. Macro-level evaluations can build on the upscaling of micro-level results or apply a separate macro-level evaluation approach. As outlined in section 3.4, both cases require plausibility and consistency checks to be carried out in relation to the data used as input and the comparison of the results at different levels. This section synthesises and reviews the key findings and experiences from the case study testing in considering micro-macro consistency.

Biodiversity (BW and HNV)

- The Farmland Bird Index (FBI) is well established as a macro indicator for biodiversity. However, it has recognised limitations in relation to micro-macro consistency. An alternative indicator, the Number of Farmland Birds Individuals (NFB), while using the raw FBI monitoring data, has been shown to be more sensitive at micro level and can be applied in other member states. The introduction of NFB strengthens the micro-macro level consistency.
- The corncrake density is another alternative indicator for use at a micro level which can be derived from the raw FBI data and which represent specific causal relationships with particular sub-measures.

Climate stability

- Two very different approaches have been tested for the assessment of the RDP impact on climate stability. Both methods rely heavily on detailed micro-level data.
- The ecological footprint is assessed by crop type and aggregated up to a macro level, Regional level results have been realised through coefficients for all the crops not analysed at micro level. Statistical tests are necessary to validate the results.
- The sectoral model is a strictly macro-level modelling method built on non-spatial micro-level data. Due to the non-spatial nature of the model, micro-macro level consistency is based on external assumptions about profit maximisation.

Landscape

- The landscape metrics method creates separate but conceptually coherent indicators for micro and macro level based on spatial statistics.
- Generally, there is consistency between micro- and macro-level assessments of impacts on landscape although such assessments can be strongly scale dependent.
- Commonly-used EU-level land cover and use classifications (e.g. CORINE, EEA) may be unsuitable for micro level because they are not at a sufficiently high spatial resolution for use in the measurement of change. For the assessment of measures for particular cultural landscapes or landscape features there may be a need for more detailed location-specific land-cover classifications.
- The use of landscape character assessments provides a basis for linking the interpretation of micro- and macro impacts on landscapes, but such datasets are not available for all countries in Europe.

Soil

- USLE can be used at a micro and macro level, however it is dependent on model input data quality; the application of USLE in both CLUE and/or InVEST requires substantial modelling effort.
- Modelling approaches (CLUE, InVEST) are macro-level approaches which are based on micro-level causal relationships with land-management practices.

Water

- The availability of monitoring data at detailed and aggregated levels from two different time periods has allowed the assessment of micro-macro consistency by aggregating the micro-level data in the same way as the macro-level monitoring data and comparing the results.

The experience of the case studies has shown that micro-level upscaling and macro-level approaches are able to provide a level of micro-macro level consistency. In the case studies, only one of these options was used; however the confidence in the micro-macro consistency of an assessment of net impacts will be enhanced if the results of micro-level upscaling and macro-level approaches can be compared for consistency.

5 Conclusions

The results from the case studies have shown that the steps of the macro-level logic model each provide an important opportunity for an evaluator to control and improve the quality of the evaluation. The selection of the unit of analysis and assessment of the indicator consistency create the foundation for a sound evaluation. The consistency between micro and macro levels is checked at this stage in the process of evaluation as well as during the final net impact assessment.

Data quality and availability are critical constraints on improvements of the RDP evaluation and the quantification of the net-impact assessment. The data-related issues are an inevitable consequence of the limited integration of monitoring strategies within environmental monitoring and RDP monitoring. Each of the monitoring strategies is designed for unique circumstances driven by public good, policy and reporting levels as well as the cost of observations. Many aspects of the RDP environmental impact evaluation can benefit from a consideration of integration of RDP monitoring with environmental monitoring programmes, in particular the causal relationships and the consistency between micro and macro level. Integration of the monitoring efforts does not always require more data, but it does require the development of a sampling strategy that can support assessments of change and policy impact.

Technological and regulatory developments towards open data access, the provision of 'big data', and its analysis, may, in coming years, contribute to addressing issues of data access.

References

Aalders, I, Morrice, J, Miller, D, Hough, R, Povellato, A, Longhitano, D, Schwarz, G (2013) *Review of macro level methods and scales.* Report to the European Commission. Deliverable D5.1, 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).

Ewert F, van Ittersum M, Heckelei T, Therond O, Bezlepkina I, Andersen E (2011) Scale changes and model linking methods for integrated assessment of agri-environmental systems. *Agriculture, Ecosystems & Environment* 142: 6-17.

Ferrier S, Drielsma M (2010) Synthesis of patterns and process in biodiversity conservation assessment: a flexible whole-landscape modelling framework. *Diversity and Distributions* 16: 386-402.

Gibson C, Ostrom E, Ahn T (2000) The concept of scale and the human dimensions of global change: a survey. *Ecological Economics* 32: 217-239.

Lukesch, R, Schuh, B (2010) *Working Paper on Approaches for assessing the impacts of the Rural Development Programmes in the context of multiple intervening factors.* Findings of a Thematic Working Group established and coordinated by The European Evaluation Network for Rural Development. pp 226.

Mauchline, A L, Mortimer, S R, Park, J R, Finn, J A, Haysom, K, Westbury, D B, Purvis, G, Louwagie, G, Northey, G, Primdahl, J, Vejre, H, Kristensen, L. S, Teilmann, K. V, Vesterager, J. P, Knickel, K, Kasperczyk, N, Balázs, K, Podmaniczky, L, Vlahos, G, Christopoulos, S, Kroger, L, Aakkula, J, Yli-Viikari, A (2012) Environmental evaluation of agri-environment schemes using participatory approaches: experiences of testing the Agri-Environmental Footprint Index. *Land Use Policy*, 29 (2). pp. 317-328.

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).

Purvis G, Louwagie G, Northey G, Mortimer S, Patk J, Mauchline A, Finn J, Primdahl J, Vejre H, Vesterager J, Knickel K, Kasperczyk N, Balazs K, Vlahos G, Christopoulos S, Peltola J (2009) Conceptual development of a harmonised method for tracking change and evaluating policy in the agri-environment: The Agri-environmental Footprint Index. *Environmental Science & Policy* 12(3): 321-337.

Reinhard A. J, Linderhof V (2013) Using spatial econometrics in impact assessment. SPARD deliverable D4.5. The Hague, LEI-Wageningen UR.

Schwarz, G, Aalders, I, Miller, D, Morrice, J (2014) *Report on monitoring and data requirements for macro level methods.* Report to the European Commission. Deliverable D5.2, ENVIEVAL project (Project Reference: 312071).

Van Ittersum M, Brouwer F (2010) *Environmental and Agricultural Modelling*. Dordrecht: Springer.