

A Practical Guide to Counterfactual Methods for Evaluating Nature-Based Solutions

Authors: Colas CHERVIER (University of Limerick); Catalina Posada Borrero (University of Montpellier)

Aim of this guide

The **BIOFIN-EU project** aims to boost **private sector investment in nature-based solutions (Nbs)** (Figure 1) by helping to unlock key **data and information bottlenecks** that currently constrain the scalability of such investments.

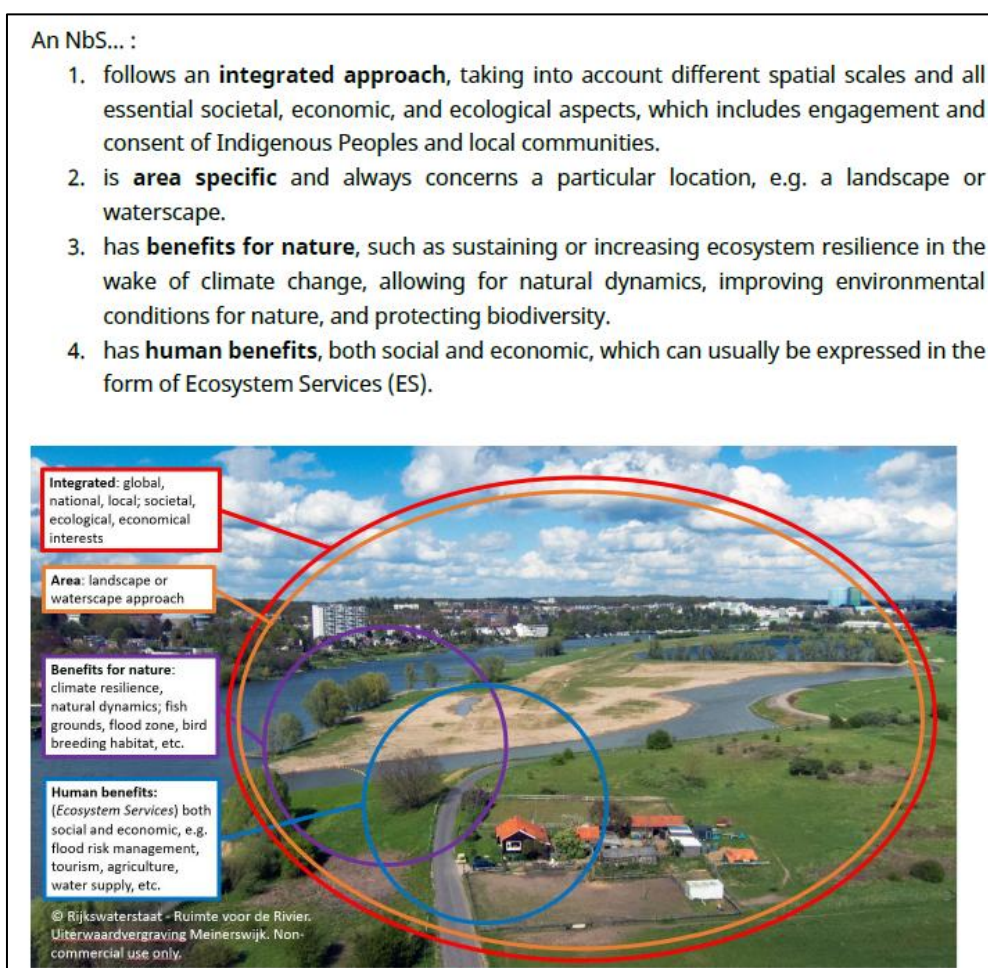


Figure 1. BIOFIN-EU’s working definition of nature-based solutions

An increasing number of private investors are placing strong emphasis on the ability to **measure and report the impact** of their investments. Doing so not only enhances the credibility of these investments but also supports progress toward broader sustainability goals, such as achieving carbon neutrality. However, in practice, limited technical capacity and insufficient data to produce credible evidence of impact remain major



barriers. These challenges can deter private investment and limit the flow of finance to otherwise promising NbS initiatives. Even when data is available, selecting the **right evaluation method** is critical, as it directly influences the credibility, acceptance, and policy relevance of impact claims.

This guide is designed to support **NbS practitioners, private investors, and the scientists and evaluation specialists** who work alongside them in addressing this key challenge: the selection and application of robust **impact evaluation methods**. It has two main objectives:

- In **Part 1**, the guide explains the **importance of using counterfactual approaches** for impact evaluation—methods that help ensure impacts are truly attributable to the intervention. It uses the case of **forest carbon credits traded on the Voluntary Carbon Market (VCM)** to highlight what can go wrong when attribution is weak, and provides an overview of the key counterfactual evaluation methods available.
- In **Part 2**, the guide offers a **step-by-step overview of how to design and implement counterfactual impact evaluations** in practice—from framing the evaluation scope to running the analysis and interpreting results. Wherever possible, the guide draws on practical **examples from deforestation-related policies, projects, and NbS interventions**.



Part 1. Why Counterfactual Methods Matter for Impact Evaluation of nature-based solutions?

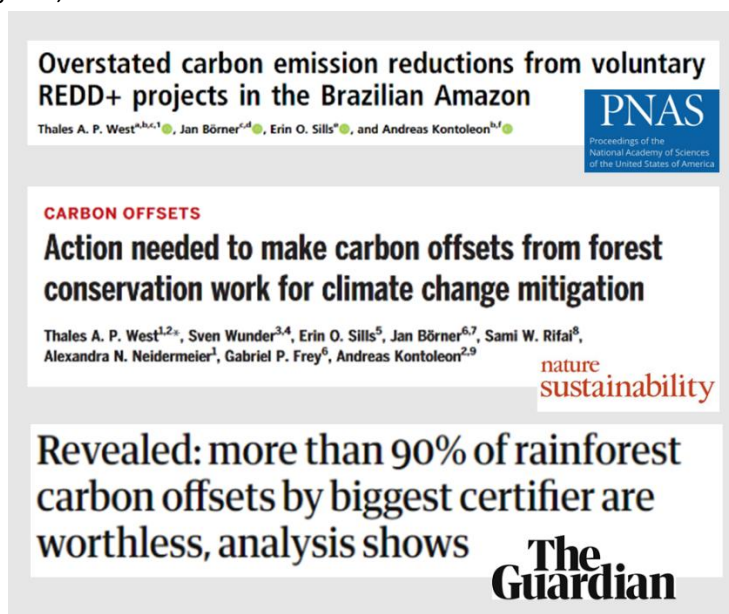
In this first part of the guide, we emphasize the importance of understanding and implementing counterfactual methods for impact evaluation of nature-based solution, using the recent controversy surrounding the forest carbon credits traded on the Voluntary Carbon Market (VCM) as a motivating example. We also provide an overview of key counterfactual methods and explain what they are designed to evaluate—laying the groundwork for applying these approaches in practice (part 2 of the guide).

The Case of the Voluntary Carbon Market (VCM)

The VCM offers a significant opportunity to finance nature-based solutions. In 2022, projects related to forestry, land use, and agriculture accounted for nearly half of all carbon credits traded on the VCM. Among these, REDD+ projects—aimed at *Reducing Emissions from Deforestation and Forest Degradation*—have emerged as a major source of carbon credits.

To be traded, forest carbon credits must undergo verification. Verification bodies—most notably Verra—develop and apply standards to certify greenhouse gas (GHG) emission reductions from REDD+ projects. A central component of these standards is the methodology used to establish baselines. These baselines are then compared to the project’s actual emissions reductions to determine its additionality and to calculate the volume of credits that can be issued.

Starting in 2021, several scientific studies and investigative articles have questioned the additionality of many verified REDD+ projects¹² (Figure 2). These critiques suggest that numerous carbon credits may be "shadow credits"—not reflecting real, additional reductions in emissions.



¹ [Action needed to make carbon offsets from forest conservation work for climate change mitigation | Science](#)

² [Revealed: more than 90% of rainforest carbon offsets by biggest certifier are worthless, analysis shows | Carbon offsetting | The Guardian](#)



Figure 2. Examples of scientific articles and media reports that have questioned the additionality of forest carbon offset projects. These publications highlight growing concerns about whether credited emission reductions from forest projects truly represent additional climate benefits beyond what would have occurred without the intervention.

This controversy has had notable repercussions for the VCM³. Following these publications, demand for forest carbon credits declined sharply, while the average price increased as buyers sought high-integrity credits—whose supply remains limited. The situation has also increased pressure on credit buyers to conduct more rigorous due diligence. Rising reputational risks, waning trust in verification standards, and growing regulatory requirements (e.g., the EU Corporate Sustainability Reporting Directive and the European Sustainability Reporting Standards) have all contributed to a heightened demand for credible, evidence-based emission reduction claims.

The Limits of Historical Baselines

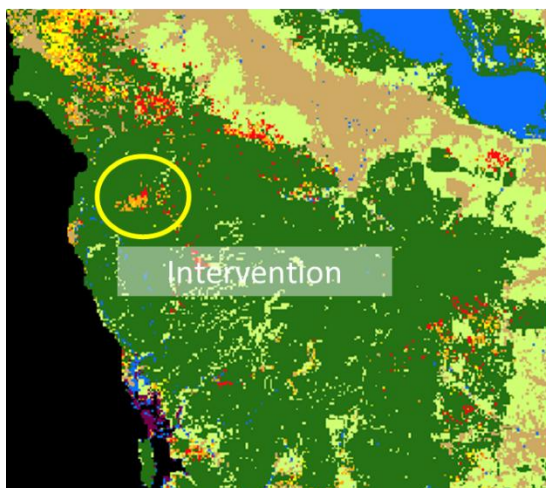
The controversy surrounding REDD+ projects largely stems from a methodological divide over how baselines should be estimated—and this is precisely where counterfactual methods become crucial. Most verification standards currently rely on historical baselines. In this approach, a project selects a baseline period and a reference area, calculates the average deforestation in the reference area over that period, and then compares this baseline with the average deforestation during the project’s crediting period. This method is essentially a *before-after* comparison of outcomes.

However, this approach fails to account for contemporaneous factors—changes in the economic, political, climatic, or institutional context—that can influence deforestation trends independently of the project. For example, an increase in palm oil prices during the project’s implementation could trigger a land rush and accelerate deforestation, regardless of project efforts. Conversely, a shift in national leadership—such as Lula’s first election in Brazil—might usher in more protective environmental policies and reduce deforestation nationwide.

When these contextual changes are not accounted for, they confound the project’s estimated impact. A simple before–after comparison may detect a reduction in deforestation, but it cannot determine whether that change was actually due to the project (Figure 3.A). As a result, such methods measure overall changes in outcomes but fail to attribute those changes to specific project activities. This is problematic because it allows projects to claim—and profit from—emission reductions that are not truly additional, but instead the result of favourable circumstances.

³ [2023 State of the Voluntary Carbon Markets Report - Ecosystem Marketplace](#)

A. Before-after comparison



B. Simple with-without comparison

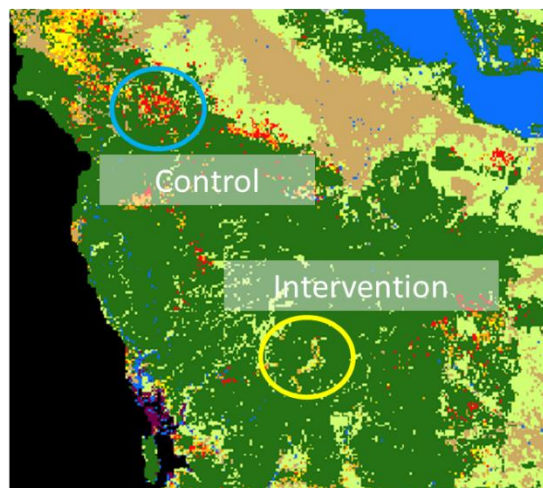


Figure 3. Limitations of before-after and simple with-without comparisons in impact evaluation. This figure uses the case of a Payments for Ecosystem Services (PES) program implemented in Cambodia’s Cardamom Mountains to illustrate why these basic comparison methods can be misleading. In Panel A, a before-after comparison misattributes deforestation (red and orange pixels) to the intervention, when in fact it was caused by the construction of a dam reservoir during the same period. In Panel B, a simple comparison between the intervention area and a randomly selected control area (blue circle) may also lead to biased conclusions, as the control area lies closer to the forest edge—where deforestation risk is inherently higher even before the intervention.

What Are Counterfactual Methods?

The most straightforward way to control for the confounding effects of contemporaneous factors is to compare the project area with a control area that is exposed to the same broad external influences—such as economic, political, or climatic changes. The assumption is that these contemporaneous factors will affect both the control and treatment areas in similar ways. If this assumption holds, then any difference in deforestation outcomes between the two can more plausibly be attributed to the project itself.

However, identifying a valid control area is far from simple. Even when control sites are located in regions subject to the same overall trends, they may still differ significantly from the project area in terms of baseline deforestation risk (Figure 3.B). For instance, selecting control villages that are closer to roads or urban centres than the intervention villages may introduce bias, as these areas are generally at higher risk of deforestation. This could lead to a misleading conclusion: if deforestation is found to be higher in the control villages than in the project villages, one might wrongly infer that the project successfully reduced deforestation. In reality, the difference could be at least partly due to the fact that the control areas were more prone to deforestation from the outset. This issue is known as selection bias, and it is the primary source of error that counterfactual methods are designed to correct.

A counterfactual represents the hypothetical scenario of what would have happened in the project area had the project not been implemented. The difference between the actual outcome and the counterfactual



outcome represents the unbiased impact of the project. However, because this alternate reality is —by definition— unobservable, it cannot be measured directly.

To approximate the counterfactual, counterfactual methods rely on comparisons between the project (or treatment) area and control areas that did not receive the intervention. To ensure that this comparison is valid—that is, that the control areas closely resemble what would have happened in the absence of the project—these methods apply scientific techniques to address selection bias.

Three counterfactual impact evaluation methods

The aim of this guide is not to delve into the econometric foundations of the methods used to address selection bias—these have been thoroughly covered in other accessible resources, such as the World Bank's *Impact Evaluation in Practice*⁴. Instead, our focus is to offer a practical, step-by-step guide for practitioners on how to apply counterfactual impact evaluation methods in real-world settings. That said, a basic understanding of how these methods work—particularly how they deal with selection bias and what their limitations are—is essential for their proper use.

We also do not cover randomized controlled trials (RCTs), which are often considered the gold standard for impact evaluation. While randomization—assigning the intervention randomly to a subset of eligible units (e.g. households or villages)—is the most effective way to construct a bias-free counterfactual, it is rarely feasible or politically and ethically acceptable in practice, especially for policies already being implemented at scale.

The remainder of this section provides a brief overview of the three most commonly used counterfactual approaches.

Matching

Matching is a widely used method to control for selection bias arising from observable pre-treatment characteristics. The core idea is to identify for each treated unit—such as a village or region participating in a program— one or more similar units that did not receive the intervention. These control "twins" are selected based on having comparable characteristics that are known to influence both the outcome and the likelihood of receiving the treatment (e.g., population density, proximity to roads, baseline deforestation rates).

By constructing a control group that mirrors the treated group in terms of these observable variables, matching aims to ensure that—on average—there are no systematic differences between the two groups before the intervention (Figure 4). The impact of the project is then estimated by comparing the average outcomes between the treated group and its matched controls.

It is important to note that matching relies on the assumption that there are no unobserved differences between the groups that also affect the outcome. If such unobserved factors exist (e.g., local governance capacity, informal institutions), the estimated impact may still be biased.

⁴ [Impact Evaluation in Practice - Second Edition](#)



Variable	Sample	Means Treated	Means Control	Diff in means
Baseline deforestation (ha)	Unmatched	-0.86	-4.20	3.34
	Matched	-0.86	-0.93	0.07
Slope (%)	Unmatched	9.67	9.18	0.49
	Matched	9.67	9.37	0.30
Area of fertile soil (ha)	Unmatched	15.21	29.50	-14.29
	Matched	15.21	14.96	0.24
Population size (nb HH)	Unmatched	92.21	175.44	-83.23
	Matched	92.21	105.00	-12.78
Accessibility (min)	Unmatched	153.23	108.72	44.52
	Matched	153.23	146.66	6.57

Figure 4. Matching. This figure, drawn from an evaluation of a PES scheme in Cambodia, shows how matching helps reduce differences in key covariates between treated (intervention) and control units. For instance, before matching, control units were, on average, located 44 minutes closer to the nearest road. After matching, this difference is reduced to just six minutes. Similar improvements in balance are observed across all covariates included in the analysis.

Synthetic control Method (SCM)

The SCM builds on the logic of matching but takes a more sophisticated approach to constructing the counterfactual. Instead of identifying a single control unit for comparison, this method creates a "synthetic" version of the treated unit by assigning optimal weights to a group of control units. These weights are chosen so that the synthetic control closely replicates both the pre-intervention characteristics and, most importantly, the outcome trends of the treated unit over a long time period (Figure 5).

This allows for a more accurate estimation of what would have happened in the absence of the intervention. A good synthetic control will have pre-intervention outcome trends that closely track those of the treated unit, lending credibility to the comparison.

The program's impact is then estimated by comparing the post-intervention outcomes of the treated unit to those of its synthetic counterpart.

As with matching, the key assumption behind the synthetic control method is that no unobserved factors affect the treated and control units differently after the intervention. If such unobservable differences exist, they can bias the estimated impact.

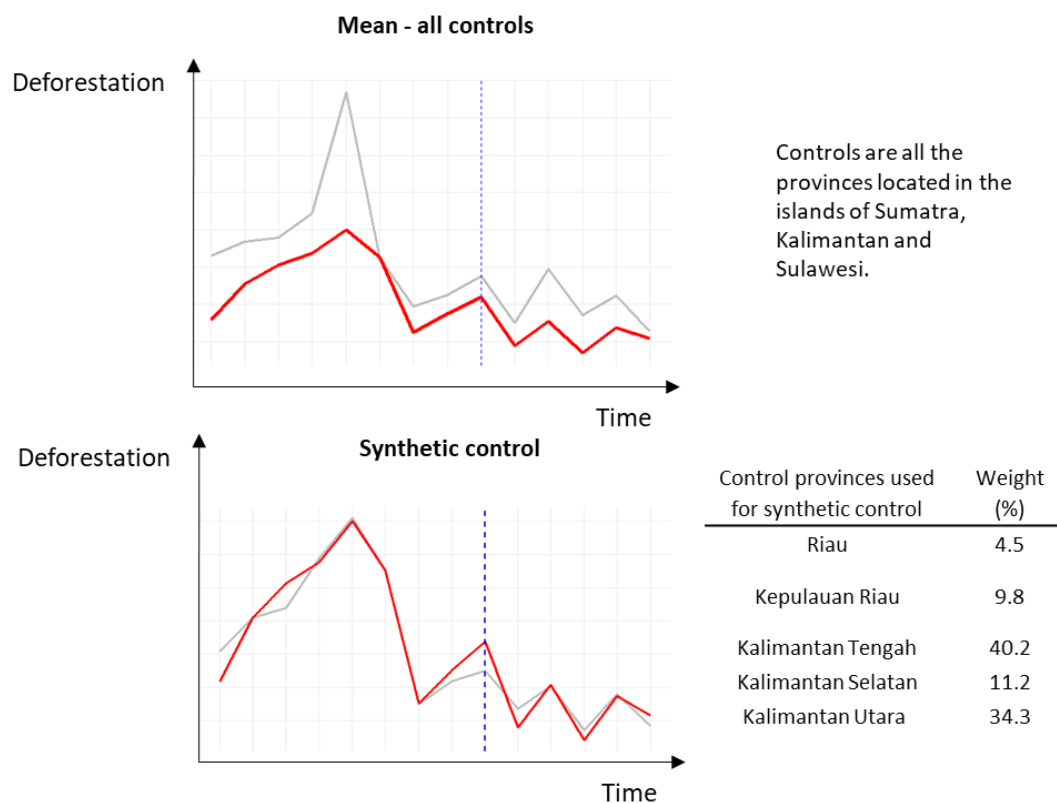


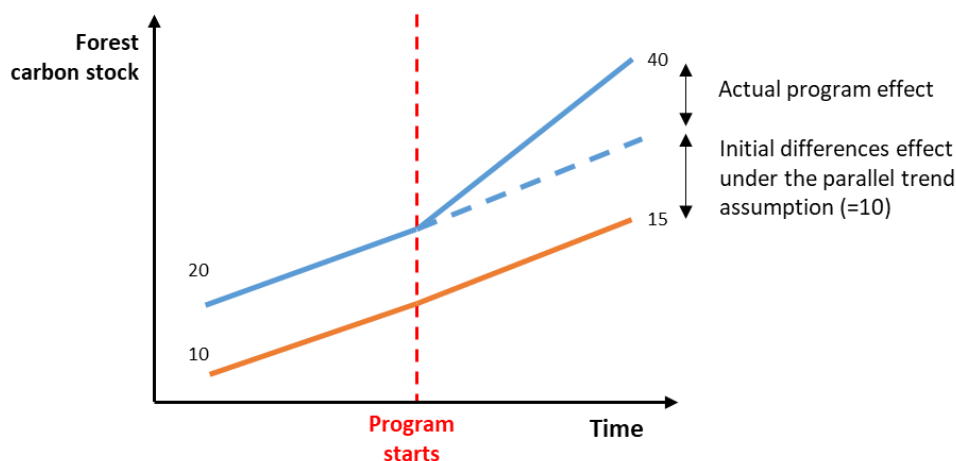
Figure 5. Synthetic Control Method. This figure, drawn from the evaluation of a forest carbon offset program implemented in the province of East Kalimantan in Indonesia, compares the evolution of the mean deforestation outcome for all control units (top panel, grey line) and for the synthetic control (bottom panel, grey line) with the treated unit (red line). The synthetic control method substantially reduced the pre-program differences between the treated and control units (program start indicated by the blue dashed line). This was achieved by constructing a weighted combination of five Indonesian provinces, with Kalimantan Tengah contributing the largest share (40%) to the synthetic control.

The Difference-in-Differences (DiD) method

The DiD method estimates a program’s impact by comparing changes in outcomes over time between a treated group and a control group. Specifically, it examines the difference in outcomes before and after the intervention for both groups, and then compares these changes to isolate the program’s effect.

This approach controls for pre-existing differences between the treated and control groups, as long as those differences remain constant over time (Figure 6). For example, if one group historically had higher deforestation due to its location, DiD will account for that, as long as that location-based difference does not change during the evaluation period.

A key assumption underlying DiD is the parallel trends assumption—the idea that, in the absence of the program, both groups would have followed similar trends over time. If this assumption is violated—for example, if a new local policy or infrastructure project affects only one group—then the DiD estimate may be biased, as it would wrongly attribute external changes to the program’s impact.



Simple before-after comparison: $40 - 20 = 20$
Simple with-without comparison: $40 - 15 = 25$
Difference-in-differences: $(40 - 15) - (20 - 10) = 15$

Figure 6. Difference-in-difference. This figure shows the hypothetical changes in forest carbon stock over time in both program and control areas. If the initial difference in forest carbon stock between the two areas—caused by location-specific factors (value of 10)—remains constant over time, even after the program begins (as shown by the dotted blue line), then the DiD method will correctly isolate the program's impact (resulting in an estimated effect of 15). In contrast, simpler approaches like before-after or with-without comparisons cannot account for this selection bias.

Summary & What's Next

This first part of the guide has briefly illustrated why it is essential for practitioners involved in the design and implementation of nature-based solutions—and seeking access to private finance—to consider counterfactual approaches as methods for impact evaluation. Assuming that private investors are increasingly interested in measuring and reporting the impact of their investments, we highlight the risks of relying on simple before-after comparisons. Such approaches may fail to demonstrate that observed ecological improvements—such as reductions in deforestation—are actually attributable to specific project interventions.

Overlooking this issue has already proven counterproductive in the case of the Voluntary Carbon Market, where weak attribution methods led to questions about credibility. The resulting controversy caused a decline in private investment supply and raised due diligence requirements—and associated costs—for investors.

We also provided a brief overview of counterfactual methods, focusing on how they address selection bias and control for contemporaneous influences, while also acknowledging their limitations. In the second part of the guide, we will turn to the practical side: outlining how these methods can be implemented step-by-step by practitioners in real-world settings.



Part 2: A step-by-step guide for the implementation of Counterfactual Methods

Step 1: Define the Specific Focus of the Evaluation

A successful impact evaluation begins with carefully framing its core focus. This step lays the foundation for the entire evaluation and ensures that the chosen methods are appropriate, relevant, and actionable for both researchers and stakeholders.

Identify the Intervention and Outcomes

Start by clearly identifying the **intervention** or group of interventions to be evaluated. These may include a specific policy, program, or nature-based solution—for example, a REDD+ initiative, a reforestation subsidy, or a sustainable agriculture incentive.

Next, define the **key outcomes** you intend to measure. These typically fall into one or both of the following categories: environmental outcomes (e.g. deforestation, forest degradation, biodiversity) and/or socio-economic outcomes (e.g. household income, wellbeing, health). Be as specific as possible. For instance, are you interested in overall forest loss, or deforestation specifically caused by fire or illegal logging?

Also consider whether to conduct **heterogeneity analysis**—an examination of whether the impact varies by factors such as geographic region, type of intervention, or contextual characteristics like land tenure systems or access to markets.

Conduct Feasibility and Relevance Checks

To help refine the focus of the evaluation, conduct several early-stage assessments:

- **Policy relevance:** Is the evaluation question aligned with the priorities of policy-makers, funders, or affected communities? Whenever possible, co-develop the research question with relevant stakeholders to strengthen ownership and ensure the findings are useful.
- **Scientific relevance:** Is the question novel and grounded in evidence? Review the academic and grey literature to understand existing knowledge gaps and how your evaluation can contribute new insights.
- **Data availability and consistency:** Check whether you have access to reliable, spatially explicit outcome data for both treatment and control areas. Very importantly, this data should also be consistent across treatment and control areas. Note that socio-economic data is often limited or unavailable outside of intervention areas, or not consistent across both areas. You'll also need accurate implementation data—preferably in the form of georeferenced shapefiles—to identify where the program took place exactly.
- **Program implementation:** Evaluate whether the intervention was implemented in practice (not just on paper) and whether it was active for a sufficient duration to plausibly generate impacts. Interventions that are poorly implemented or too recent may not be suitable for evaluation.





Develop a Theory of Change

Creating a **Theory of Change** is a critical step in refining your evaluation scope. A Theory of Change clarifies how the intervention is expected to produce outcomes, and helps establish whether there is a logical reason to expect an impact at all. If there is little likelihood that the intervention will produce a measurable effect, conducting a rigorous evaluation may not be justified.

A well-crafted Theory of Change also strengthens the evaluation design by helping to focus the evaluation questions (e.g., which specific outcomes are being assessed? What heterogeneity analysis to conduct?), identify relevant control variables, and inform the choice of counterfactual method.

Below and in Figure 7 are the core components to include:

- **Expected Outcomes and Impacts.** Define what the intervention aims to achieve in both the short term (outcomes) and the long term (impacts). For example, an expected outcome might be a reduction in illegal logging, while the long-term impact could be improved forest ecosystem health or climate mitigation.
- **Selection Mechanism.** Reflect on why the intervention was implemented in certain locations or among certain groups. Was it due to higher deforestation risk, political feasibility, or logistical access? Understanding this helps identify and adjust for potential selection bias, which is essential for accurate impact evaluation.
- **Project-Specific Interventions.** Clearly identify the concrete actions undertaken as part of the intervention. Most projects or policies consist of multiple components—for instance, community training, enforcement patrols, or financial incentives—that should be listed and understood individually.
- **Causal Pathways.** Explain the mechanism of change: how do the intervention's activities lead to the intended outcomes? Identify the intermediate steps and enabling conditions needed for change to occur. Incorporating intermediate outcomes into the analysis helps capture key nuances along the causal pathway, offering a more comprehensive understanding of how and why the intervention generates (or fails to generate) its intended impacts.
- **Unintended Effects and leakages:** Anticipate potential unintended consequences of the intervention, such as environmental degradation or decrease in wellbeing. A common issue is *leakage*, which occurs when efforts to reduce deforestation in one area displace the harmful activity, and deforestation, to another. For example, stricter enforcement in protected forests may push logging or agricultural expansion into nearby unprotected lands. Including such risks in the Theory of Change helps ensure a more realistic and responsible assessment of the intervention's overall impact.
- **Accounting for External Influences.** Consider external factors that could also affect the outcomes, such as changes in commodity prices, rainfall patterns, market demand, or national-level policy reforms. Recognizing these helps in designing a counterfactual that isolates the project's true contribution.
- **Heterogeneity Hypotheses.** Anticipate how the impact might vary across different settings, population groups, or implementation approaches. For instance, impacts might differ between remote and accessible areas, or between communities with differing governance structures.

Importantly, the Theory of Change should be developed **in collaboration with key stakeholders**. Their participation helps ensure that the causal logic is realistic, context-specific, and that the evaluation addresses questions that are relevant and meaningful to those involved. Stakeholder engagement also increases the likelihood that the evaluation results will be used to inform future decisions.

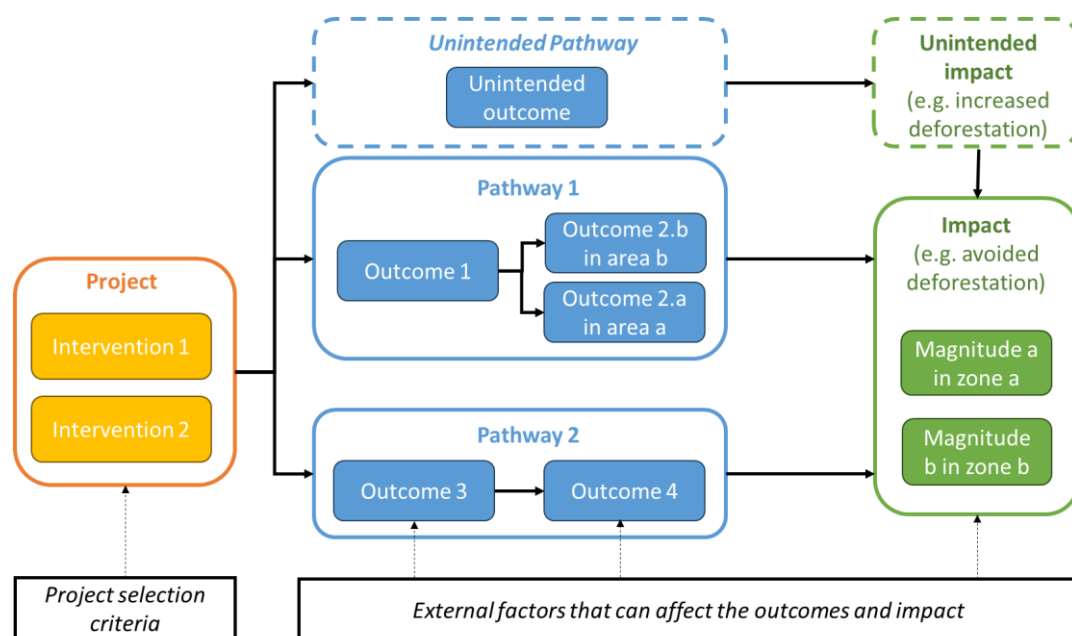


Figure 7. Generic Theory of Change for Impact Evaluation. This figure outlines the core components of a theory of change, including interventions, intermediate outcomes, and long-term impacts. It also emphasizes the importance of accounting for impact heterogeneity, unintended effects, selection mechanisms for project locations, and external contextual factors that may influence outcomes.

Step 2: Identifying the Building Blocks for the Impact Evaluation Method

Once the focus of the evaluation has been clearly defined, the next step is to identify the **key building blocks** needed to structure your impact evaluation design. This includes defining the **unit of analysis** and identifying the **treated and control units** required to conduct a credible counterfactual analysis.

Define the Unit of Analysis

The **unit of analysis** is the level at which data will be compiled and comparisons made between treated and control units. In counterfactual impact evaluations of nature-based solutions (NbS), commonly used units of analysis include:

- Households
- Grid cells (e.g., 1 km² spatial units)
- Jurisdictional or administrative units (e.g., villages, municipalities, provinces)

Choosing the appropriate unit of analysis is a critical step, as it affects the quality and feasibility of the evaluation. The following considerations can help guide the choice:

- **Availability of outcome data.** Socio-economic data are often available at the household or administrative unit level, while environmental outcomes like deforestation are typically measured using spatially explicit data (e.g., raster .tif files) at the grid cell level (Figure 8).
- **Level of intervention implementation.** While it's generally recommended to align the unit of analysis with the level at which the intervention is implemented, there may be good reasons to select smaller units. For example, if there is reason to believe that the intervention's effectiveness varies across individuals or areas, smaller units allow for more detailed heterogeneity analysis.
- **Number of potential control units.** A sufficient number of control units is needed to construct a credible counterfactual. Methods such as matching and the SCM perform best when there is a large pool untreated units to draw comparisons from. This often supports the use of smaller units of analysis, which increase the sample size.
- **Computational feasibility.** On the other hand, working with smaller units significantly increases the number of observations and computational demands. If analytical or technical capacity is limited, using larger units of analysis may be more practical. In some cases, aggregating smaller units into larger ones can also help satisfy the parallel trends assumption.

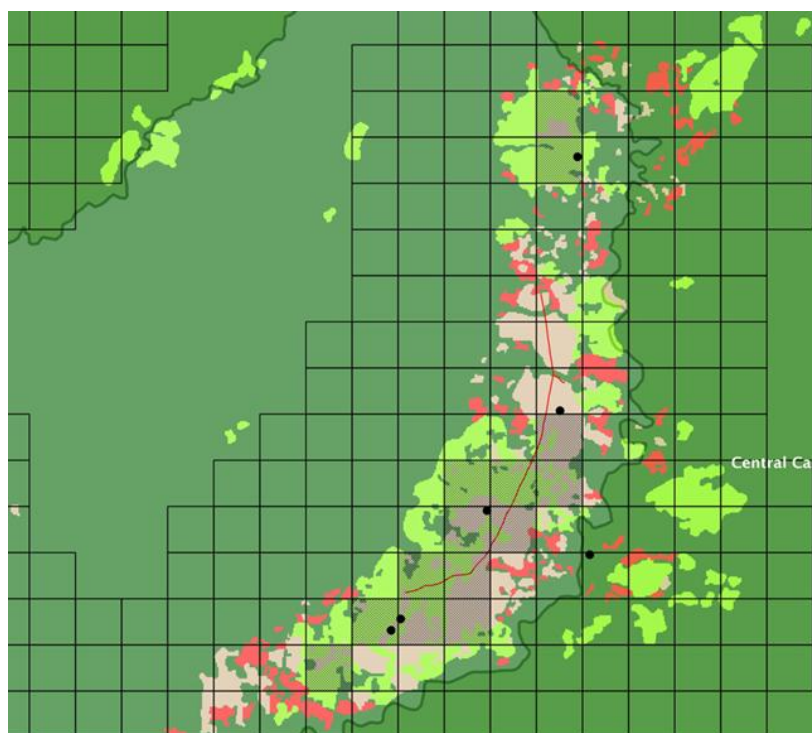


Figure 8. *Example of Grid Cells as Units of Analysis.* This figure shows how a 1x1 km grid was used to assess the impact of a Cambodian PES scheme on deforestation. A 5 km buffer was drawn around each participating community, and the buffered area was divided into grid cells. This approach leverages the raster format of the



outcome variable (e.g., satellite-based forest loss at 30 m resolution) and enables the creation of a large number of spatial units for analysis.

Identify the Treated Units

Treated units are those spatial or administrative units that were exposed to the intervention and where the project's impact is expected to occur. This step is often straightforward but requires careful documentation. Ideally, evaluators should secure access to a spatially explicit dataset (e.g., shapefiles) indicating the exact location and boundaries of the intervention areas.

It is important to note that many nature-based policies and programs—such as community forestry or conservation schemes—are often implemented in **non-contiguous areas**. For example, a national social forestry program might be rolled out in scattered communities across a country, depending on eligibility criteria or administrative capacity.

When identifying treated units, be sure to document the following additional elements. These details will be particularly valuable for analyzing heterogeneity of effects, as outlined in the Theory of Change:

- **Timing of Treatment.** Clearly document the start and end dates of the intervention for each treated unit. This is particularly important if the intervention was rolled out in phases or implemented at different times across locations. If timing varies, it must be accounted for in the evaluation design.
- **Type of Intervention.** Under a single policy or program, the form of intervention may vary by location. For instance, in the case of protected areas, different sites may carry different legal statuses or enforcement regimes (e.g. strict nature reserves vs. multiple-use areas). Similarly, community forests may differ in focus depending on the implementing partner. A development NGO may emphasize livelihood generation, while a conservation NGO may prioritize biodiversity protection. These distinctions should be carefully documented, as they may lead to different impact pathways or magnitudes—and are highly relevant for heterogeneity analysis.
- **Level of Implementation.** Even within the same intervention type, the intensity or quality of implementation may vary across sites. For example, social forestry programs may receive differing levels of support from government agencies or NGOs, affecting local implementation capacity and outcomes. Where possible, evaluators should gather information on variation in implementation quality, which can later be used to explain differences in effectiveness across treated units using heterogeneous treatment analysis.

Select Appropriate Control Units and selection criteria

This step focuses on identifying a **pool of eligible control units** that can serve as the foundation for estimating the counterfactual— that is, what would have occurred in the absence of the intervention.

The identification of potential control units should begin with a **broad but contextually appropriate geographic scope**, such as an entire country, a specific state or province, or an ecologically defined area like a biome. The selected area should be large enough to provide a sufficient number of potential control units, but also homogeneous enough to ensure comparability with treated units. For instance, it is not advisable to combine dry deciduous forests with tropical moist forests when evaluating the impact of an intervention on deforestation, as the ecological dynamics and drivers of land-use change may differ substantially. Similarly, selecting control units that span across national borders may introduce serious challenges due to differences



in institutional frameworks, political environments, and policy implementation capacity, all of which can independently influence the outcomes of interest.

Within this broader area, a more **refined pool of control units** should be selected. Some counterfactual methods, such as matching or the SCM, provides tools to select a subset of the most suitable control units from within this initial group. In contrast, Difference-in-Differences (DiD) approaches may require the evaluator to manually screen or pre-match potential controls.

At this stage, defining appropriate **selection criteria** is critical. These include:

- Control units should share **similar pre-treatment characteristics** with treated units, particularly those that are likely to influence post-treatment outcome trends. For example, control units should be drawn from communities that are similar in terms of ethnic composition, socio-economic conditions, and institutional capacity, or from forests that have comparable ecological characteristics and face similar deforestation pressures.
- Additionally, selected control units should be located in **areas where the intervention could plausibly have been implemented**, or where it might be implemented in the future. This underscores the importance of understanding the criteria used to select intervention areas in the first place.
- It is crucial to ensure that selected control units are not indirectly influenced by the intervention, hence **avoiding contamination of control units**. While neighbouring units may appear to be good comparators, they can be affected by spillover effects—for example, reduced deforestation pressure due to increased enforcement or community engagement in adjacent treated areas, or conversely, a displacement of deforestation activities into nearby control units. Such contamination can bias the results, leading to either an underestimation or overestimation of the intervention's effect, and ultimately compromise the validity of the impact evaluation. If these spillover effects are anticipated in the Theory of Change, it is essential to identify and exclude control units that are likely to be exposed to them.
- It is also essential to ensure that control units are not subject to **other policies or interventions** that could influence the outcome of interest. Overlapping initiatives make it difficult to isolate the effect of the intervention under study. For instance, when assessing the impact of a protected area network, it would be important to exclude any control units that overlap with mining concessions, logging permits, or other natural resource management programs such as social forestry. Failing to do so risks conflating the effects of multiple policies, thereby biasing the evaluation.

The identification of control units and the selection criteria should involve **consultation with local experts and stakeholders** who are familiar with regional dynamics and the implementation of the intervention. This can be done informally through expert interviews or more formally through participatory sessions, such as those used in co-developing a Theory of Change.

Finally, a common constraint in identifying suitable control units that should be considered at this stage is the **availability and quality of data**. A credible comparison requires detailed pre-treatment data to characterize potential control units and confirm their eligibility. Reliable spatial, ecological, and socio-economic data are essential, and evaluators should prioritize consistency across both treated and control areas. Where **local or regional data sources** are more accurate than global datasets, they should be preferred to support a more robust evaluation design.



Step 3. Design the Estimation Strategy

Once treated and control units have been identified, the next step is to design a sound estimation strategy—that is, to choose how the project’s impact will be measured. This involves selecting the most appropriate method to estimate the counterfactual impact, to assess the robustness of results and to measure impact heterogeneity. The main goal is to ensure that any observed changes in outcomes can credibly be attributed to the intervention rather than to chance or unrelated external factors.

This step requires **more advanced econometric skills**, and it may be advisable to **seek guidance from experts**.

Selecting the Right family of Method

Choosing between the three main families of methods—Matching, DiD, and the SCM—depends on three key data-related criteria (Figure 9).

First, both DiD and SCM require **pre-intervention data on the outcome variable**, such as annual deforestation rates or forest degradation indices. This data is essential for verifying the assumption that treated and control units followed similar trends before the intervention. Unfortunately, in many real-world evaluations, outcome data is only collected for the treated areas—for instance, when a project conducts a baseline survey exclusively in its target communities. In such cases, Matching is the most viable method, provided there is enough reliable information available on pre-intervention covariates to build a credible counterfactual.

Second, SCM is particularly well-suited for cases where there is **only one—or a small number of—treated units**. A common example is a policy or program implemented in a single administrative area, such as a pilot community forestry initiative launched in just one province. However, recent methodological extensions, such as Generalized Synthetic Control (GSC), have made it possible to accommodate multiple treated units within a similar analytical framework. By contrast, Matching and DiD approaches typically perform better when a larger number of treated units is available, as this improves the statistical power and reliability of the estimates. Regardless of the method chosen, all counterfactual approaches tend to yield more robust results when there is a relatively larger pool of untreated units compared to treated ones.

Third, Matching and SCM rely on detailed **pre-intervention data for covariates**—variables that affect both the likelihood of receiving the intervention and the outcome of interest. These covariates must be measured consistently across treated and control units. However, in practice, such data is often missing, incomplete, or recorded in incompatible formats for control units, making valid comparisons challenging. In these situations, DiD becomes a more viable alternative, as it does not require covariate balance but instead relies on the parallel trends assumption. If this assumption does not hold due to observable differences between groups, it is possible to condition on covariates by including them in the DiD regression model to adjust for these imbalances and help recover parallel trends. Even when the assumption appears valid, incorporating time-varying covariates can further strengthen the DiD analysis by accounting for factors that influence outcome dynamics over time.



	Matching	DiD	SCM
Pre-intervention data on covariates in both control and treated units available	Necessary	Not necessary	Necessary
Large number of treated units available	Necessary	Necessary	Not necessary
Pre-intervention data for the outcome in both control and treated units available	Not necessary	Necessary	Necessary

Figure 9. Criteria for Selecting an Appropriate Counterfactual Impact Evaluation Method. This diagram outlines the key data requirements and contextual factors that guide the choice between Matching, DiD and SCM. The tree is intended to help evaluators select the most suitable family of methods based on the availability of outcome and covariate data, the number of treated units, and the structure of the intervention.

Stay Updated with Methodological Developments

Once the main family of counterfactual methods has been selected, the next step is to choose the **specific estimation method** to be used. The field of counterfactual impact evaluation is evolving rapidly, with new estimators and techniques continuously emerging to improve and refine existing methods. As such, conducting a **targeted literature review** is essential—not only to identify the most up-to-date and robust estimators, but also to select the one that best fits the specific context of your study.

For example, **recent advances in DiD** have addressed some of the method’s longstanding limitations, including how to deal with **staggered treatment timing** and **heterogeneous treatment effects** across units. These improvements are particularly relevant in policy evaluations where programs are implemented at different times across regions or target populations that may respond differently to the intervention.

Conduct Robustness and Sensitivity Tests

To ensure the credibility of your findings, it is essential to conduct **robustness and sensitivity tests**. These checks help determine whether the estimated impacts reflect a true effect of the intervention or are simply the result of modelling choices or data quirks.

Sensitivity analysis involves testing whether the results remain consistent under reasonable variations in model specifications. For example, does the estimated impact of a reforestation subsidy hold if you use a slightly different outcome variable—such as canopy cover instead of forest loss—or the same outcome variable sourced from a different but reliable dataset. Similarly, adjusting the set of covariates included in the model can help assess whether results are robust to changes in variable selection. Another useful approach is to apply alternative estimation techniques within the same methodological family. For instance, comparing results obtained using Mahalanobis distance matching with those from propensity score matching can reveal whether findings depend too heavily on a specific matching algorithm. These checks are essential for building confidence that the observed impact is not an artifact of modelling choices.

Placebo tests are another powerful tool for assessing robustness. These typically involve estimating the intervention’s impact in a period before it was actually implemented, or applying the model to a unit that resembles a treated unit (picked in the pool of control units for example) but was never treated. If a significant effect is found in these placebo scenarios, it suggests that the main model may be overfitting the data or capturing random noise rather than a true causal impact. Placebo tests are particularly common in studies using DID and SCM, where they help assess whether the estimated effect is unique to the treated unit or could be replicated in untreated units by chance.



Explore Heterogeneity in Impact

Finally, a strong estimation strategy should go beyond measuring the average treatment effect and **explore heterogeneity**—how the impact of the intervention varies under different conditions. Two main sources of heterogeneity should be considered: **variation in context characteristics** (such as ecological, institutional, or socio-economic conditions) and **variation in intervention characteristics** (such as the nature of the implementing agency or the degree of enforcement).

Exploring heterogeneity not only improves the scientific rigor of the evaluation but also generates practical insights for decision-makers. It helps answer critical questions like: *Where should interventions be prioritized?* and *What types of approaches are most effective under which conditions?*

In the literature, two main approaches are commonly used to assess heterogeneity. The first involves **dividing treated units into subgroups** based on a key contextual or implementation variable—for example, comparing outcomes in high-risk versus low-risk areas—and estimating impacts separately. This approach can highlight meaningful differences in policy effectiveness across different settings.

The second approach models **how the impact varies along a continuous scale**, such as distance to infrastructure, elevation, baseline forest loss rates, or intervention characteristics like enforcement intensity. This technique provides more granular insights into how effects evolve across a spectrum of conditions, but it typically requires a larger sample size and a wide distribution of the contextual variable. It also tends to be more technically demanding, requiring more advanced statistical modelling skills.

Regardless of the approach, it is essential that the heterogeneity analysis is grounded in the **Theory of Change** developed earlier. Doing so ensures that variations in impact are interpreted within a logical causal framework and helps avoid drawing conclusions from spurious or coincidental patterns in the data.

Step 4. Running the analysis

Once the estimation strategy has been chosen and the required data collected, it's time to implement the analysis. This phase includes preparing the datasets, running the estimation using the appropriate tools, and interpreting the results in a way that is both statistically sound and relevant to real-world decision-making.

Data collection

If the evaluation relies on **secondary data**, much of the effort will go into preparing and processing spatially explicit datasets such as shapefiles, raster files (e.g. satellite imagery), and administrative records like census data or land-use classifications. These datasets often require cleaning, standardization, and, in many cases, spatial merging or reclassification, which may demand specific GIS and spatial data management skills. For instance, evaluating a deforestation reduction project may involve integrating satellite-derived forest loss data with the boundaries of project and control areas, along with spatial layers for rainfall, roads, and population density.

In addition to using secondary sources, impact evaluations—particularly those focused on socio-economic outcomes—often require the collection of **primary data**. This typically involves a **household or community-level survey** conducted in both treated and control areas. The survey must be carefully designed to capture not only the relevant outcome variables (such as income, food security, or fuelwood collection) but also pre-treatment covariates that influence both outcomes and the likelihood of receiving the intervention. Ideally, a



baseline survey is conducted in both groups before the intervention begins. However, when this is not possible, post-intervention surveys should still include retrospective questions to collect data on pre-treatment characteristics such as household size, land tenure status, or education levels.

Database creation

Once data collection is complete, the next step is to **construct the working database**. The format of the dataset will depend on the counterfactual method chosen (Figure 10). For instance, **Matching methods** typically require data in a **"wide" format**, where each observation unit (e.g. village, household, or grid cell) occupies a single row and each variable—such as baseline income, forest cover, or proximity to roads—is a separate column. In contrast, **DiD** and the **SCM** require data in a **"long" format**, where each unit has multiple rows corresponding to different time periods (e.g. annual deforestation rates from 2005 to 2020), and time is recorded as a separate variable.

While these general formatting rules are useful, it is essential to refer to the **documentation of the specific R or Stata package** you are using, as each may have its own precise requirements for variable names, structure, or inputs. Following these instructions carefully is critical for avoiding errors during analysis and ensuring that results are valid.

Creating these datasets often involves merging spatial, temporal, and survey-based data—sometimes from multiple sources—and checking for internal consistency and completeness. This step can be complex and may require iterative cleaning and reformatting.

Unit id	Treatment (0/1)	Covariate 1 pre-treatment	Covariate 2 pre-treatment
1	0	810	25
2	0	937	45
3	0	1077	36
4	0	1448	2
5	0	890	8
6	0	1035	12
7	1	1053	23
8	1	1030	65
9	1	840	85

Wide-format database for matching

Unit id	Year	Treatment (0/1)	Covariate (time-varying)	outcome
1	2001	0	810	25
1	2002	0	937	45
1	2003	0	1077	36
2	2001	0	1448	2
2	2002	1	890	8
2	2003	1	1035	12
3	2001	0	1053	23
3	2002	1	1030	65
3	2003	1	840	85

Long-format database for DiD & SCM

Figure 10. Database formats: Wide vs. Long. This figure illustrates the structural differences between wide and long database formats. In the wide format (typically used for Matching), each unit (e.g., village or household) appears in a single row, with separate columns for each variable or time period. In the long format (used for DiD and SCM), each unit appears in multiple rows—one per time period—with a dedicated column for the time variable.

Data analysis

Running this type of analysis is rarely a straightforward process—it often involves **multiple rounds of trial and error** to arrive at a satisfactory model. A "satisfactory" model is one that meets the **quality criteria specific to the chosen method**. In Matching, for example, this means achieving **good covariate balance**—that is, ensuring there are no statistically significant differences in the means of key variables between the treated and control groups after matching. For DiD and SCM, the main criterion is whether the **pre-treatment trends** in the outcome variable are closely mirrored between the treated unit and its comparison group or synthetic counterpart.



If these conditions are not met, it may be necessary to **adjust the specification**—for instance, by trying different combinations of covariates (particularly relevant for Matching and SCM), or by revising or expanding the pool of control units, which is often the case in DiD and SCM applications where counterfactuals are weak. That said, if earlier steps—such as defining the unit of analysis, selecting appropriate control units, and designing the estimation strategy—have been conducted carefully and systematically, the need for major adjustments at this stage should be minimal. Still, some level of iteration is to be expected and should be considered a normal part of the analytical process.

Interpretation of results

Finally, once results are obtained, the focus shifts to **interpretation**. Counterfactual methods provide a **quantitative estimate of the intervention's impact**, the results of significance testing, and—if heterogeneity analysis was conducted—some insight into impact pathways. However, they cannot explain everything. Reviewing findings in light of **existing quantitative and qualitative studies** can help assess consistency with past evidence and better understand the mechanisms at play. Moreover, it is critical to **interpret findings in collaboration with domain experts and local stakeholders**, who can contextualize the results, explain anomalies, and ground the evaluation in the socio-political realities of the region.

In some cases, complementing quantitative results with **qualitative work**, such as interviews, focus groups, or field visits, may help provide a fuller understanding of the mechanisms behind observed changes.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.