Restrictions of Empirical Policy Analyses: The Example of the Evaluation of Rural Development Policies

Anne Margarian*

***Johann Heinrich von Thuenen-Institute, Institute of Rural Studies, Bundesallee 50, D-38116 Braunschweig, e-mail: anne.margarian@vti.bund.de**

Paper prepared for presentation at the 118th seminar of the EAAE
(European Association of Agricultural Economists),
'Rural development: governance, policy design and delivery'
Ljubljana, Slovenia, August 25-27, 2010

Abstract

The present paper asks under what circumstances a standardisation of evaluations would be feasible in order to enable a comprehensible aggregation of results for the European administration. We argue that in the complex environment of rural development the adequate definition of system boundaries is a precondition for the successful application of empirical methods and the identification of causal effects. If macro effects and self-enforcing effects are important, the objects of inquiry have to be defined on a higher observational level. In this case, the statistical identification may not be possible because there might be hardly any comparable ("counterfactual") observations. We conclude that evaluators need definite theoretical guidance in order to define consistently their field of inquiry. Only then, the goal of comparable and aggregable quantified results might be achievable to a certain degree.

Keywords: Evaluation, Complex Systems, Causal Inference, Counterfactual Approach

JEL Codes: O22, Q18, R58, C51

1 Introduction: The aim of quantitative impact assessment

The Rural Development Policies of the member states of the European Union are co-financed by the European Community "[s]ince the objective of this Regulation, namely rural development, cannot be achieved sufficiently by the Member States given the links between it and the other instruments of the common agricultural policy, the extent of the disparities between the various rural areas and the limits on the financial resources of the Member States in an enlarged Union". The Council assumes that common objectives "can therefore be better achieved at Community level through the multiannual guarantee of Community finance and by concentrating it on its priorities […]" (Council Regulation (EC) No 1698/2005). In order to justify the flow of financial funds from different administrative levels, member states have to give account for their achievements to the Commission, the Commission has to give account to the European Parliament and all political institutions have to give account to the European public. Nevertheless, there is an overwhelming amount of information to be processed from the evaluation of the effectiveness of numerous measures summarised in the Rural Development Plans[1]. Within these plans in accordance with the requirements of the European Commission, "Strengths, Weaknesses, Opportunities and Threats" of the "rural areas concerned" are analysed ("SWOT analysis") (DG AGRI, 2006). Thereby, the heterogeneity of the areas covered by each single plan is described.

In the face of the resulting overwhelming amount of data and information accumulating at all administrative and political levels, in the last decade the European Commission spent high effort to implement institutions for an ongoing and all-embracing evaluation in the field of rural-development-policies. After the summarization, analysis and interpretation of data by evaluators for all areas covered by distinctive Rural-Development-Programs (RDPs), further processing of the delivered information by the central European Agencies is necessary. Due to the reports' diversity and heterogeneity, this task remains an excessive demand, which makes it impossible

[1] There are 15 Rural Development Plans for Germany, four for the United Kingdom, three for Portugal, six for France, seven for Italy, five for Spain, two for Belgium, two for Finland and one for each of the other member states. (http://ec.europa.eu/agriculture/rurdev/countries/index_en.htm)

for central European Agencies to take all the information delivered into consideration in the decision-making processes. Consequently, increasing attempts have been made to standardise the evaluation of RDPs.

Fields of standardisation are the timing and the determination of indicators for evaluations. With respect to methods applied, the less ambitious aim is "to establish good practice"[2]. Concerning the timing of evaluations, they are to be delivered as "Midterm-" and "Ex-post-" Evaluations. Apparent deficits of this time frame, which is mainly motivated by administrative reasons, lead to the creation of the concept of the "ongoing" evaluation for 2007-2013. For the same programming-period, far-reaching commitments in the standardisation of indicators have been made. For each measure potentially included in RDPs, as well as for the programs as wholes, specific indicators have been determined that are to be used by evaluators[3]. Within the European Evaluation Network for Rural Development, experts still work on the identification of methods that could be proposed as "best practice" approaches. The question of interest is, whether the institutionalised evaluation with the described approach of standardisation will be able to deliver the desired condensed and comprehensive information.

The first problem is that of different weights for different aims in different (regional) contexts. The single additionally created job for example will be valued differently in boom-regions with a shortage of qualified labour than in remote areas. A simple aggregation of changes in the respective indicators that are caused by the implemented measures could therefore create a misleading picture.[4] This paper, though, concentrates on the second, even more fundamental problem: the consistent and reliable identification of the quantified effects of interventions. Only if every evaluator of each measure everywhere in the European Union accomplishes the task to quantify reliably the measures' effects, the danger of comparing and aggregating apples and oranges is prevented. As the official Evaluation Guidelines for Rural Development Programmes (Guidelines) put it: "Impacts will be identified as net-contributions of each single measure to achieving a programme's objectives" (Guidance Note B: 10). This paper analysis, what the restrictions and preconditions would be if standardised evaluations with comparable results were to be guaranteed. Thereby we ask for the theoretical and practical possibility of standardised evaluations with comparable results capable of being totalled.

In order to make the point we start out by a discussion of the current situation of the evaluation of Rural development programmes (chapter 2). Described is the approach to evaluation as it is propagated by the European Commission today (section 2.1) and the advanced approaches that are based on the paradigm of the counterfactual (section 2.2). Chapter 3 prepares the ground for a critical reflection on the application of these approaches to the evaluation of RDPs. Therefore, in section 3.1 we introduce briefly the idea of complex systems. Then the idea

---

[2] This aim may be extracted from the description of the purpose of the European Evaluation Network for Rural Development, which has been implemented by the European Commission's Directorate-General for Agriculture and Rural Development. HTML: http://ec.europa.eu/agriculture/rurdev/eval/network/whatwedo_en.htm (Last access on March, the 9th 2010).

[3] On http://ec.europa.eu/agriculture/rurdev/eval/index_en.htm (Last access on March, the 9th 2010) in Annex 3 guidance documents concerning baseline indicators, output indicators, result and impact indicators are offered.

[4] There are two principal ways to overcome this problem. One is the consistent application of a general cost-benefit-calculation. Nevertheless, such an approach is extremely demanding. On the one side, the value of each political goal would have to be quantified on a common, probably monetary, basis. Secondly, in the evaluation all positive and negative impacts of each measure would have to be identified and quantified as well. Another, more pragmatic way to overcome the problem of regionally differing weights of aims, is the definition of zones, which are characterised by specific problems. The aims and observed changes in indicators could then be aggregated for the respective zones separately.

that rural development takes place in a complex system is defended (section 3.2). Chapter 4 discusses the problem of identifying causal effects in complex systems. Therefore, central problems of the counterfactual approach in complex environments are discussed (section 4.1). Consequently the role of a priori causal knowledge is stressed (section 4.2) before finally an approach is proposed that allows for the necessary "economising on causal knowledge" (section 4.3). The paper concludes with consequences for the evaluation system of RDPs in chapter 5.

2 Standard approaches towards the evaluation of Political Programmes

2.1 The evaluation approach of the European Commission

It is the aim of the European Commission to be informed by evaluations about the impacts of measures and programmes. This is to be achieved with the help of "impact indicators". "These refer to the benefits of the programme both at the level of the intervention but also more generally in the programme area" (Evaluation Guidelines, Guidance Note B: 5). The Guidelines mention the necessity of the construction of a counterfactual situation, but they remain rather vague: "As evaluation looks at change over time, the establishment of the counterfactual is a central issue for all evaluations. In this context the 'base-line indicators', established by the SWOT analysis and ex-ante evaluation at the time of programming, need to be mentioned. The base-line indicators are an important reference point for the evaluation of impacts of single measures and programmes as a whole". No further elaboration exists on how these base-line indicators are defined and how these "reference points" help in the identification of program effects. "The evidence for impacts shall be provided by indicators which refer to the benefits of the programme beyond the immediate effects in its direct beneficiaries both at the level of the intervention but also more generally in the programme area. They are linked to the wider objectives of the programme. They are normally expressed in "net" terms, which means subtracting effects that cannot be attributed to the intervention (e.g. double counting, deadweight), and taking into account indirect effects (displacement and multipliers)" (Lukesch and Schuh, 2010).

We might most clearly point to the problems of this concept by a discussion of the term "net effect" and an analysis of the opposed concept of a "gross effect". If an effect consists of the impact of a cause, what should be the possible meaning of a "gross effect"? One explanation we find is "Beneficiaries' statements are called "gross effects" (including bias) whilst the evaluation team's estimate is called a "net effect" (corrected from bias)"[5]. One could deduce that gross effects have to be understood as the observed correlation between intervention and changes in impact indicators. The advice is, to primarily calculate the gross effect and then to isolate the net effect by subtracting the potential confounding influences. In order to follow this advice, though, we not only have to know the potentially influential confounding factors but we also have to know the magnitude of each single one of them. Therefore, under this concept, "[i]t will be difficult, if not impossible to fully close the attribution gap. The system under observation – the impact of policy interventions on rural areas – is too complex to be grasped comprehensively" (Lukesch and Schuh, 2010).

It seems to be recognised that there are fundamental problems in the identification of effects of certain measures but a constructive and offensive approach on how to handle this problem is not proposed. Rather, evaluators are advised to do something anyway: "As a consequence, the chosen methodology will in many cases be a "second-best solution", based on a trade-off between what should be done and what can be done" (Lukesch and Schuh, 2010: 20). The opinion seems to prevail that evaluation approaches the truth if it does the best it can.

[5] http://ec.europa.eu/europeaid/evaluation/methodology/methods/mth_att_en.htm

Nevertheless, in the presence of ignorance, the results of statistical inferences may be as misleading as a naïve judgement based on observed correlations as has been shown by "Simpson's paradox" (Simpson, 1951; see also Pearl, 1999). Due to this effect, the inclusion or deletion of a single variable could even reverse the sign of an estimated coefficient. If the estimation is not stratified properly, the effect of the bias might be fatal, no matter how advanced the applied methodology is.

Often "a mix of quantitative and qualitative approaches and methods within a coherent overall architecture" (Lukesch and Schuh, 2010) is proposed in order to reduce the danger of systematic biases caused by certain methodological approaches. Nevertheless, it is probable that biases of different approaches have the same direction. A selection bias for example, will be as prevalent in statistical analysis as it is in the experience of participants or experts, who might be interviewed, because they, too, observe mainly correlations instead of causations. Some more advanced statistical approaches to the identification of causal effects are presented in the following.

2.2 Causal models and the concept of the counterfactual

"Advanced" empirical (statistical) approaches towards evaluation rest on the fundamental idea of the counterfactual. The comparison of the outcomes of two otherwise identical situations, one with intervention and one without, is used in order to identify the effect of the intervention by differencing. The counterfactual approach thereby tries to avoid the problem to quantify each single causal relation within the development under scrutiny. Since we are principally unable to observe the two states with and without intervention on the same subject simultaneously, experiments under controlled conditions are applied regularly in the natural sciences. Such experiments mean that "the individuals or material investigated, the nature of the treatments or manipulations under study and the measurement procedures used are all selected, in their important features at least, by the investigator" (Cox and Reid, 2000 cited by Morgan and Winship, 2007). Due to practical and ethical problems, we observe controlled experiments much less frequently in the social, economic or political sphere. One exemplary example of a thorough field-experiment in the controlled application of a social program is described in the literature concerning the Mexican program PROGRESA[6] (especially Behrman and Todd, 1999 and Skoufias, 2005). The careful discussion of the results of this field-experiment shows that even experiments in the context of the implementation of social programmes are by no means comparable with controlled experiments in a laboratory environment. The main problems in the field occur in the range of the necessary verification of random assignments of treatments.

Actually, if we were able to conduct controlled experiments in the narrowest meaning of the concept, we might in certain cases be able to isolate effects of policies without any further knowledge of causal relations that contribute to changes in the objects under investigation. Nevertheless, in policy evaluation we will usually only be able to construct a *hypothetical* "experiment that could ideally be used to capture the causal effect of interest" (Angrist and Pischke, 2009). These hypothetical experiments will help to assess the quasi-experiments that economists and other social scientists in the vast majority of cases have to deal with. Quasi-experiments in the definition of Cook and Campbell (1979) "have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred" (cited by Morgan and Winship, 2007). Therefore, we need an "identification strategy" that allows for the usage of the corresponding observational data in order to approximate a real experiment (Angrist and Pischke, 2009). The "identification

[6] http://www.ifpri.org/search/publications?keys=PROGRESA

strategy" depends on the right identification of those causal influences that have to be controlled and those that may be ignored. Obviously, in order to meet this decision, we have to know about potential causal influences. The advantage of the quasi-experimental approach is that we do not have to quantify all of them. In the quasi-experimental approaches, the counterfactual situation is not controlled in an experiment. Rather, the counterfactual is designed by means of an ex-post controlling of influential variables.

The matching approach most obviously reflects the idea of the counterfactual. In the matching process, those observations that have a maximal similarity with respect to the relevant characteristics are identified. Thereby the confounding variables are controlled ex post, i.e., after the intervention, in the comparison of "matched" treated and non-treated. In propensity-score matching, a score is estimated that quantifies the probability of participation of each subject. This score serves as a measure of comparability. It can be shown theoretically, that those subjects that are similar in their probability of participation are also similar in possibly influential variables with respect to the relevant development under intervention. Nevertheless, in this approach only observable and known differences may be controlled.

In order to overcome this problem, the difference-in-difference approach has often been applied, sometimes in combination with the statistical matching approach (e.g. Pufahl and Weiss, 2009). The underlying idea is simple: under the assumption that initial differences between subjects impact upon the level but not on the course of development, only differences in the development of the subjects of interest are observed. Under the assumption that all subjects are exposed to the same changes in environmental conditions and that there are no endogenous differences in individual trajectories, the observed differences in the development may then be ascribed to the intervention. The approach has been further refined by the fixed effect panel estimation. Here, different exogenous dynamics may be controlled in the panel-estimation approach as long as exogenous time-variable influences are observable and known. With respect to endogenous developments that are potentially correlated with the intervention a simple solution for statistical controlling does not exist (Morgan and Winship, 2007: 254ff). Morgan and Winship also stress "the deep connections between regression and matching as complementary forms of a more general conditioning estimation strategy" (p. 165). Consequently, we will not discuss the concrete choice of a certain statistical methodology and the strength and weaknesses of different approaches here. More generally, we focus on the fundamental problem of conditioning or stratification of observations. We discuss this problem in the light of the complex-system paradigm in the following.

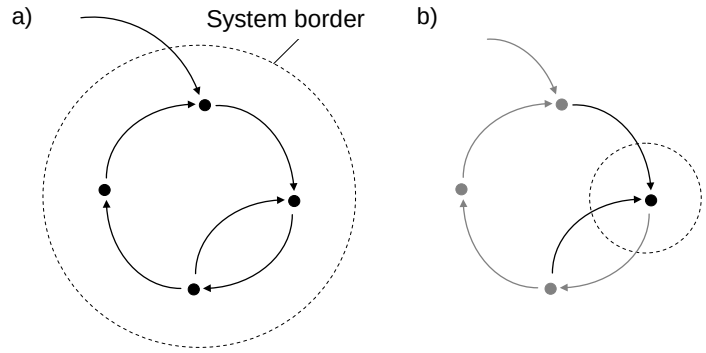
3 The complexity of Rural Development

3.1 Complex Systems

The counterfactual approach allows under certain circumstances for the isolation and identification of "relationships among observable variables [that] remain invariant when the values of those variables change relative to our immediate observation" (Pearl, 2000b). It is thereby an approach that follows the analytic or "reductionist" (Mitchell, 2009) paradigm of science. The traditional analytical paradigm of science has been developed and applied successfully in the 18th and 19th century. The analytic approach is characterised as "separating something into component parts or constituent elements" (Merriam-Webster's Online dictionary). It is supposed that understanding the parts suffices in order to reconstruct the functioning of the whole. Nevertheless, this analytic approach is not adequate for the inquiry of complex systems (Mitchell, 2009). We might call a system complex if it consists of a large

number of different interrelated elements[7] or if there are self-enforcing and/or self-inhibiting relations between the system's elements. Complexity in this sense comprises "problems which involve dealing simultaneously with a sizable number of factors which are interrelated into an organic whole. They are all, in the language here proposed, problems of organized complexity" (Weaver, 1948).

Thinking in complex systems has to be different, because complex systems are characterised by non-linear relations and emergent phenomena. Emergence is defined as "the arising of novel and coherent structures, patterns and properties during the process of self-organization in complex systems" (Corning, 2002). Non-linear relations are often caused by reflexivity that is, by the property of the complex system to react on its own situation. This characteristic is especially prevalent in social systems, which have therefore been characterised by the sociologist Luhmann (2004) as "self-referential". Consequently, its own state determines in part the development of complex systems endogenously. For these reasons, the scientific inquiry of complex systems rather focuses on the relation between elements of the system than on the elements themselves. In analytical approaches, in contrast each single influence is conceptualised as exogenous. This is achieved by a sufficiently narrow bordering of the object of inquiry as described in figure 1 b as in contrast to figure 1 a.



Source: Alfeld and Graham, 1976, p.62, modified

Figure 1: System borders and specific analyses in the presence of feedback loops

The analytical approach and the approach of complex systems are not alternatives but they complement one another. In system-theoretic inquiries, too, the selected boundaries of the object under inquiry never comprise all existing feedback loops (Alfeld and Graham, 1976: 62). The central question than is that concerning the adequate cut-off border with respect to the environment. As Pearl (2000a: 349f) says: "If you wish to include the entire universe in the model, causality disappears because interventions disappear – the manipulator and the manipulated loose their distinction." The corresponding empirical problem is that with very large units of comparison the number of units to be compared is getting to small in a statistical sense.

Figure 1 illustrates within a causal-loop-diagram the problem of identifying the adequate level of inquiry. The adequate border of the system under scrutiny includes the relevant feed-back-loops, because only then will potential self-reinforcing or self-inhibiting feed-back-loops and the indirect effects they mitigate be captured adequately. Arrows, which have no ancestors themselves within the system-border may be treated as exogenous and can be controlled for example within fixed-effects-

[7] A large number of identical elements lead to "disorganized complexity" (Weaver, 1948). This disorganized complexity is exactly what statistical approaches are used for. Consequently, we do not deal with this kind of complexity here.

panel models by the inclusion of corresponding time-varying variables. If on the other hand the system-border is defined narrowly like in figure 1b, potential indirect effects are not captured.

3.2 The complex policy-context of rural development

The economy is a social system. We mentioned in the last section that social systems are often perceived as complex in the defined manner. In economics, the perfect market assumption has commonly bypassed the threat of complexity towards the analytical approach. In the given context of Rural Development, there are several reasons for dropping the perfect market assumption that guarantees the stable functional relations as they are assumed in numerous modelling approaches toward policy assessment. The first reason is the actuality of interventions themselves. In market economies, interventions will usually be justified by specific malfunctions of markets as the following statement clarifies: "In particular there is a presumption that the market mechanism should be left to function freely unless there is some identified 'market failure' or public distributional objective. A market failure is defined as 'an imperfection in the market mechanism that prevents the achievement of economic efficiency' (HM Treasury, 2003, p.103) and economic efficiency in turn has a specific technical definition" (ADAS, 2004). Since our market-based societies are generally built on similar assumptions concerning potential welfare-effects of policies, we should expect to meet imperfect markets wherever policy-interventions are observed. Otherwise, the intervention should be questioned on a theoretical basis.

Actually, basic aims of the Rural Development Policies are the reduction of disparities in regional development and the creation of similar living-conditions in all parts of the European Union.[8] Disparities in regional developments nowadays are not being explained anymore solely based on differences in the natural endowment of a region nor solely in reference to transport-costs. One of the reasons for the enhancement of theoretic explanations of regional disparities has been the observed ineffectiveness of policies that these older theories had guided (compare Puga, 2002). The New Economic Geography (NEG) and other dynamic models of cumulative causation have acknowledged "the significance of increasing returns for spatial differentiation" (Fingleton, 2007). The NEG helped to understand the "self-reinforcing character of spatial concentration" (Fujita et al., 1999). It did so by taking into account the possibility of market imperfections like positive external effect of scale, non-tradable inputs and monopolistic competition. Positive exogenous effects of scale may for example be caused by the so called "home market effect", that is, the endogenously caused growth of the market in growing regions, as well as by interlinkages between firms (Puga, 2002). On the other side, given imperfect competitive industries, an endogenous specialisation of regions that is independent from initial factor endowment may evolve. We observe self-enforcing processes as they are characteristic for complex systems. Baldwin and Martin (2004) show that the combination of models of NEG with models of regional growth causes specific characteristics of regional development like "growth-linked circular causality" and "growth poles and sinks".

The corresponding theories should constitute a pillar within the scientific foundation of Rural Development Policies, given the current level of knowledge. Nevertheless, this means that non-linear relations between relevant factors, policy and development have to be taken into

[8] The article 130a of the Treaty that establishes the European Community states that the Union "shall aim at reducing disparities between the levels of development of the various regions and the backwardness of the least favoured regions, including rural areas". In the Council Regulation (EC) No 1698/2005 on support for rural development by the European Agricultural Fund for Rural Development it is stated that the rural development policy should "take into account the general objectives for economic and social cohesion policy set out in the Treaty and contribute to their achievement […]".

account as well as the possibility of heterogeneous, possibly adverse policy-effects on the regional level and intended or non-intended macro-effects.[9] The response of an agent in one region might be different from the response of an agent in another region towards a policy change and indirect effects might be positive in one region and negative in the other.

There are types of models and applications that run under the perfect market assumption. Nevertheless, our theoretic and empirical knowledge does not suffice presently in order to formulate similar comprehensive models that take into account the known deviations from the perfect market assumption in regional development. The natural consequence is to rely on empirical approaches. In an empirical sense, referring to the preceding discussion, the effect of the intervention depends on the context or, in a statistical sense, on intermitting variables. What is even more important in our context is that controlling the environment in the assessment of net-effects might become crucial, since otherwise effects of policies might be over- or underestimated due to omitted variable bias and selection-effects. The question then is under what circumstances in this complex setting the identification of net effects will be possible in the empirical approach. In the following chapter, we discuss the critical assumptions of the standard evaluation approaches in the light of the paradigm of complex systems and draw conclusions.

4 Evaluation in complex systems

4.1 Problems of the counterfactual design in complex systems

In the foregone chapter it has been clarified that rural development probably may be conceptualised as a complex system in that manifold different elements interact with each other, while their interaction itself depends on each region's state. In a system with many interrelated entities, the estimation of direct effects on the treated might not suffice in order to assess an intervention's impacts. Heckman et al. (1999) clearly identify the problem: "A full evaluation entails an enumeration of all outcomes of interest for all persons both in the current state of the world and in all the alternative states of interest, and a mechanism for valuing the outcomes in the different states." The authors also stress that in the traditional evaluation literature usually the effect of the measures on active participants is identified (the "Direct Effects"), while the indirect effects on participants and non-participants is ignored.

Ignoring indirect effects on non-treated is fatal due to the reliance of the counterfactual approach on the "stable unit treatment value assumption" (SUTVA) (Rubin, 1980). SUTVA requires "that the potential outcomes of individuals be unaffected by potential changes in the treatment exposures of other individuals" (Morgan and Winship, 2007). The principal idea is easily understood: if a treatment affects non-participants, and the effect on the treated is isolated by the comparison of participants and non-participants, then an over- or underestimation results, depending on whether there are positive or negative spillover effects.

Additionally, due to incomplete knowledge, lacking indicators and the unresolved question of endogenous development, heterogeneous effects have to be considered. If the heterogeneity is large, we may question the relevance of the mean treatment effect that is usually calculated in regression- as well as in matching approaches. "In this case […], there is a distinction between the parameter 'the mean effect of treatment on the treated' and the 'mean effect of randomly

[9] Potential heterogeneity of effects of regional policies given assumptions of NEG have been analysed by Baldwin and Okubo (2006): "Taking production subsidies as an example, we show that regional policies tend to attract the least productive firms since they have the lowest opportunity cost of leaving the agglomerated region (or not moving there in the first place)." On the other hand, if subsidies were sufficiently high, even more productive firms might be attracted and positive agglomeration-effects might finally arise, contributing to further endogenous regional growth. In regions with a more positive economic environment, the effects of even low subsidies might have larger positive regional impacts (probably on cost of some other region).

assigning a person with characteristics X into the program'" (Heckman et al., 1999). Moreover, in the presence of unobserved reasons for heterogeneity, the estimates of treatment effects might even be biased (Elwert and Winship, 2010).

Even more puzzling than the problem of indirect effects upon different groups are indirect effects that evolve due to dynamic endogenous processes. As Pearl (2000a) explains: "Likewise, an economist is concerned with the effect of taxation in a given economical context, characterized by various economical indicators, which (again) will be affected by taxation if applied. Such context-specific causal effects cannot be computed by simulating an intervention in a static Bayesian network, because the context itself varies with the intervention […]." As a possible reaction, the question may be formulated in such a way, i.e. the system under scrutiny may be bordered so narrow in time and space, that macro effects are simply not observed (compare figure 1). With such a design, though, only very specific questions concerning short-term consequences of interventions under very specific circumstances may be answered.

The internal validity (Campbell und Stanley, 1963) of corresponding analyses may be given. Internal validity depends on the answer to the question: „Did the experimental treatment make a difference in this specific experiment?" (ibid; Chen und Rossi, 1987). Nevertheless, in order to assess the economic meaning of the result, the scenario, i.e. the influences that have been fixed in figure 1b, has to be described accurately. As discussed in the introduction, a simple aggregation of changes within different environments would be analogous to the summation of apples and oranges. This means that with very restrictive scenarios the external validity of results of empirical investigations of intervention may be questionable. External validity depends on the answer on the question: „To what populations, setting, treatment variables and measurement variables can this effect be generalized?" (Campbell und Stanley, 1963; Chen und Rossi, 1987). The concentration on very small sub-population jeopardises the relevance of analyses: "Holding things constant by using homogeneous populations as subjects may severely undermine external validity as to render experimental results worthless for policy purposes" (Chen und Rossi, 1987).

Moreover, due to the violation of SUTVA with endogenous dynamics and resulting macro-effects, the effect may not be isolated by a comparison of comparable people from the same region. Individuals in different regions, though otherwise identical, would not be comparable due to the effect of regional characteristics on the measure's effects: "Social stratification results in almost perfect separation in propensity scores across subclasses, which renders treatment effect estimates meaningless" (Oakes, 2004). The treatment effect would not be identifiable, because that "neighborhood effects are endogenous renders any efforts to control for 'selection' between neighbourhood variability attributable to only chance and measurement error" (Oakes, 2004). Manski (1995) discusses the "reflection problem" that occurs if the composition of the population in a region has an impact upon the population's further development (neighbourhood effect in a narrow sense) and if this expected development or the composition of the population itself are correlated with the participation of individuals in the intervention. In this case, one cannot answer whether the development causes the participation or the participation causes development.

Due to the problems with indirect effects in the counterfactual design on the micro-level "the randomized community trial is canonical design for neighbourhood effect studies in particular, and social epidemiology more generally" (Oakes, 2004). If a higher observational level is chosen, though, comparable environments have to be identified. These might not exist at all. Additionally, in order to identify comparable higher-level units, the causal mechanisms relating the environment and the other peoples' influence to interventions' effects have to be known.

We conclude that some strong assumptions have regularly to be made in order to justify the application of certain statistical methodologies. These assumptions usually run contrary to the characteristics of complex (social) systems. Therefore, in order to isolate intervention effects

empirically, we have to rely on context-specific causal model. The justification of assumptions of a specific model has to be based on a theoretical model, upon which the scientific community has agreed. Such a model may be subject to future fundamental revisions, but it represents the present common paradigm of science in the sense of Kuhn (1962). The questions of the following chapter are how much causal knowledge is necessary in order to construct and defend reliably identification strategies and how we can tell whether an effect may be identified empirically, given the existence of a commonly agreed upon causal model.

4.2 Supplementing the counterfactual approach with theory-guided causal inference

It became apparent in the foregone discussion that there are some principle restrictions to the application of statistical approaches. That is the case if in the presence of endogenous dynamics the system boundaries have to be set wide and if on this high level of observation too few (if any) comparison groups exist. In every case, though, the identification of adequate boundaries of the objects under inquiry is a necessary precondition for the successful application of any empirical method to the identification of intervention-effects. Setting these boundaries as wide as necessary and as narrow as possible is a precondition of efficient evaluations. In this section an approach is proposed that helps identifying relevant and negligible variables and thereby supports the economising on causal knowledge. Moreover, it is going to be shown that additional causal knowledge may be necessary in correctly bordered systems in order to identify causal relations. Many authors have acknowledged the necessity of theoretic guidance, most prominently by the Nobel laureate Heckman, who emphasizes "the importance of understanding the decision rule and its relationship to measured outcomes in formulating an evaluation model" (Heckman et al., 1999). In the following, we rely on the graph-theoretic modelling of causal relations as it has been developed by Judea Pearl (2000a).

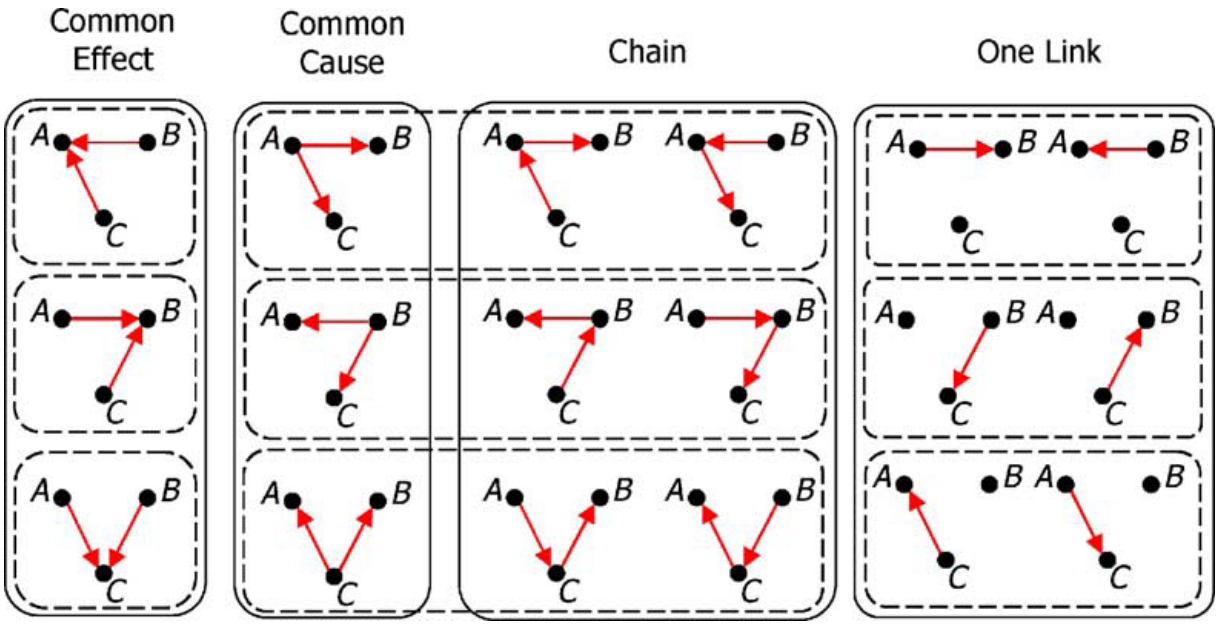
This graphical approach to causal models offers the possibility to test under what circumstances specific causal models may be identified empirically. Causal relations are presented within directed graphs (Pearl, 2000a) as in figure 2. "Nodes represent continuous or discrete state variables of a system and arrows represent direct causal relations, pointing from causes to effects" (Steyvers et al., 2003). Variables that act as causes are often labelled as "parents" of variables that represent effects. Given that the cause- and effect structure is known, a joint probability distribution over the state-variables is defined:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)).$$

A given probability distribution may be represented by many possible variants with respect to the ordering of state-variables. Under the Markov condition "the state of any variable is probabilistically independent of its non-descendants given the states of its parents" (Steyvers et al., 2003). Additional to knowledge of adequate system boundaries and relevant variables, the Markov condition is necessary for causal inference: "Causal inference exploits this principle to reason backward from observed patterns of data to the causal structure(s) most likely to have generated that data" (Steyvers et al., 2003). In the presence of macro-effect and endogenous dynamics the Markov condition does not hold. As outlined above in this case the level of observation would have to be changed in order to allow for the analytic approach of inferred causation (chapter 3).

The Markov condition is no sufficient condition, though, because different network-structures may be Markov equivalent, "meaning that they will in general produce data with the same set of conditional independence and dependence relationships. […]Without further knowledge, observational data alone provide no way to distinguish Markov-equivalent structures, such as the common-cause and chain networks" (Steyvers et al., 2003). Figure 2 shows all possible types of three-node causal networks with one or two arrows and the Markov equivalent types. We concentrate on the first line of figure 2 and interpret B as a farmer's

investments, A as the farmer's liquidity and C as the farmer's long-term strategy. If we had to choose among the three hypotheses summarised in the common cause- and the chain-structure we had no possibility to clarify on an empirical basis, whether the liquidity determines investments and strategies directly. The alternatives are Markov equivalent: investments could determine liquidity with liquidity determining the strategy or vice versa. Nevertheless, if we had the possibility to determine the farm's liquidity exogenously, we could analyse the effect of this intervention and differentiate between the hypotheses: either both, investments and strategies, should be affected, or only one of them, depending on the hypothesis confirmed by the test. Pearl (2000a) has discussed the vital relevance of interventions in this sense for causal inference. Nevertheless, which hypotheses might be tested, depends on the variables that are manipulated by the intervention (Steyvers et al., 2003).



Remark: Solid lines group together networks of the same topological type. Dashed lines delineate Markov equivalence classes.
Source: Steyvers et al., 2003
Figure 2: All possible types of three-node causal networks with one or two arrows

Often the intervention itself is not independent of influencing variables, though. In these cases it is necessary to condition on (or stratify by) the confounder. The confounder is the variable that affects the treatment and the outcome simultaneously. Since the social systems under investigation usually have no natural border, one may find additional potentially confounding variables for each confounding variable. Therefore, the identification of the leanest system possible for the question at hand is a necessity for practical evaluation. This, too, may be illustrated and accomplished with the graphical approach outlined above.

4.3 Economising on causal knowledge

Under certain conditions, conditioning on observables secures independence of the outcome variable under scrutiny from the confounders; "From a graphical perspective, the result of such a modelling strategy is to generate simplified subgraphs […]" (Morgan and Winship, 2007). Pearl (2000a) calls this independence-creating conditioning "d-separation". A d-separation along Z occurs if X is independent of Y given Z in the distribution represented by the respective Markovian graph: "A path p is said to be d-separated (or blocked) by a set of nodes Z iff:

(i) p contains a chain i→ j → k or a fork i ← j → k such that the middle node j is in Z, or

(ii) p contains an inverted fork $i \rightarrow j \leftarrow k$ such that neither the middle node j nor any of its descendants (in G) are in Z.

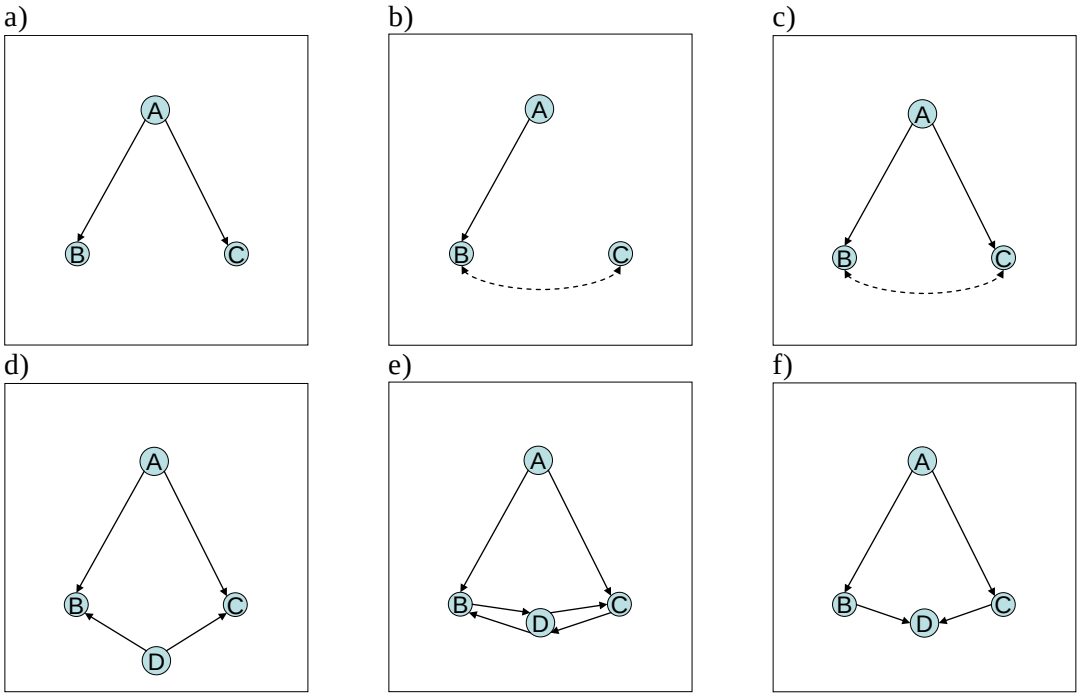
If X, Y, and Z are three disjoint subsets of nodes in a DAG G, then Z is said to d-separate X from Y, denoted $(X \perp\!\!\!\perp Y \mid Z)_G$ iff Z d-separates every path from a node in X to a node in Y" (Pearl and Dechter, 1996). Here, "DAG" means a "directed acyclical graph".

Given this concept of "blocking" paths, an admissible set of factors that allows for the identification of the causal relation under scrutiny may be identified by the "back-door criterion: "A set S is admissible (or 'sufficient') for adjustment if two conditions hold:

(1) No element of S is a descendant of X

(2) The elements of S 'block' all 'back-door' paths from X to Y, namely all paths that end with an arrow pointing to X" (Pearl, 2000a).

Here, X is the cause and Y the effect under scrutiny. Consider the graphs depicted in figure 3. Dotted two-sided arrows represent correlations between variables that are caused by common causes, which are not specified by the theory at hand or simply omitted within the graph. In figure 3 we could for example interpret A as the treatment and B and C as the relevant state variable of the treated (connection to A) or non-treated (no connection to A).



Source: own figure

Figure 3: Identifiability in very simple graphs

Referring to figure 3, in a), if we control the treatment A, we clearly "block" each path from A to B and C. Therefore, we are able to estimate the effect of A on either one of them consistently. In b), the same holds true, because the development of the non-treated C is independent of the treatment itself. It is no descendant of A. There is no "back-door path" from A to B. In c), B and C both participate in the treatment. At the same time, the development of B is affected by the development of C (and vice versa). Therefore, the effect of A on B may not be estimated consistently, because $A \rightarrow C \rightarrow B$ represents an unblocked "back -door path". Accordingly, in order to identify the effect of A on B one would have to condition additionally on C. This necessity might be cumbersome, once there are more than two agents potentially participating. In that case, it might be helpful to capitalise on theoretic knowledge about the

causes of the correlation between B and C as depicted in figure 3 d) by the introduction of D with its directed arcs.

If, for example, we knew that the development of the two agents B and C are correlated due to scarcity of land, then D could represent the current land price. Conditioning on the land-price would block the back-door paths for all potential participants that belong to the same market. The inclusion of this higher-level variable could therefore help to identify the model with reduced information. Nevertheless, if not only the price affects the farms' developments but also the farm's development (demand for land) affects the price as in figure 3 e) the graph is no DAG any more. It is not identifiable. In this case, the only solution is the complete fallback on a higher observational level, where the development of B, C and D is treated as the description of one single scenario that is to be compared with a second scenario.

The necessity of theoretical knowledge becomes even more obvious if one considers that an arbitrary treatment of variables in the hope that "more helps more" may even create biases in before non-confounded problems. Consider figure 3 f. Here, the situation of the treated affects some other variable, for example the situation of a supplying firm, while it is itself largely unaffected by the supplier's situation. An observer may recognise the correlation and decide to include D in the estimation. While B and C given A were initially uncorrelated, B (C) is now dependent on C (B) given A and D given the causal graph in figure 3 f. Controlling for D, the suppliers situation opened an additional back-door (A → C (B) → D → B (C)) accord ing to the definitions of d-separation given above. The estimated treatment effect is going to be biased.

5 Consequences for the evaluation of RDPs

We conclude that in order to enable comprehensible and comparable studies, the objects under scrutiny should be bordered as broad as necessary but as narrow as possible. This proceeding enables generalisation on the one hand, while on the other "economising" on causal relations restricts the complexity of the analytical process itself.

Consequently, theoretical knowledge and knowledge-based classifications of the objects under inquiry are important prerequisites of any empirical evaluation. Therefore, the proposal of "best-practice methods" alone will not suffice in order to guarantee comparable evaluations and results. Additionally, clear intervention-logics that serve as theoretical foundations and all embracing definitions of adequate boundaries of all objects under scrutiny for each single measure are necessary.

The European Commission strives for the implementation of an ongoing, all-embracing, high-quality institutionalised evaluation that allows for comparability of results across reports of different origins and their aggregation on a European level. Considering the identified preconditions and the current state of knowledge, we doubt the achievability of this goal in the near and medium-term future. Based on the difficulties identified, we propose an alternative approach to evaluation. The assessment of impacts, in contrast to the monitoring of outputs and certain results, should be separated from institutionalised evaluation. Scientific experts should conduct impact assessments as in-depth-studies in major projects.

Literature

ADAS Consulting Ltd (2004). Entry to and Exit from Farming in the United Kingdom (RMP 2037): Executive Summary. Prepared for the Department for Environment, Food and Rural Affairs (United Kingdom). http://www.defra.gov.uk/evidence/economics/foodfarm/reports/documents/Entry.pdf

Alfeld, L. E. and Graham, A. K. (1976). *Introduction to Urban Dynamics*. Cambridge (Massachusetts): MIT Press.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton (New Jersey): Princeton University Press.

Baldwin, R.E. and Martin, P. (2004). Agglomeration and Regional Growth. In Henderson, J.V. and Thisse, J.-F. (eds.), *Handbook of Regional and Urban Economics, Vol. 4, Cities and Geography*. Amsterdam: Elsevier, 2671-2712.

Baldwin, R.E. and Okubo, T. (2006). Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting. *Journal of Economic Geography* 6 (3): 323-346.

Behrman, J. R. and Todd, P. E. (1999). Randomness in the experimental samples of PROGRESA (education, health, and nutrition program). International Food Policy Research Institute (IFPRI). Washington, DC: IFPRI http://www.ifpri.org/sites/default/files/publications/behrmantodd_random.pdf

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Chen, H.-T. and Rossi, P. H. (1987). The Theory-Driven Approach to Validity. *Evaluation and Program Planning* 10 (1): 95-103.

Cook, T. D. and Campbell D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally.

Corning, P. A. (2002). The Re-Emergence of "Emergence": A Venerable Concept in Search of a Theory. *Complexity* 7 (6): 18–30.

Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments*. Boca Raton: Chapman & Hall/CRC.

DG AGRI, Directorate General for Agriculture and Rural Development (2006). *Handbook on Common Monitoring and Evaluation Framework: Guidance document. Annex 1C: Ex-ante Evaluation Guidelines including SEA.* http://ec.europa.eu/agriculture/rurdev/eval/index_en.htm

Elwert, F. and Winship, Ch. (2010). Effect Heterogeneity and Bias in Main-Effects-Only Regression Models. In Dechter, R., Geffner, H., and Halpern, J. Y. (eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. UK: College Publications, 327-336.

Fingleton, B. (2007). New economic geography: some preliminaries. In Fingleton, B. (ed.), *New Directions in Economic Geography*. Cheltenham: Edward Elgar, 11-52.

Fujita, M., Krugman, P. and Venables, A. (1999). *The Spatial Economy: Cities, Regions and International Trade*. Cambridge, MA: MIT Press.

Heckman, J.J., LaLonde, R. J. and Smith, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter O. and Card, D. (eds.), *Handbook of Labor Economics*, Vol. IIIA, Chapter 31. Amsterdam: Elsevier, 1865-2085.

His Majesty's Treasury, 2003. *The Green Book: Appraisal and Evaluation in Central Government*. London: TSO.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Luhmann, N. (2004). Einführung in die Systemtheorie. Baecker, D. (ed.). Heidelberg: Carl-Auer.

Lukesch, R. and Schuh, B. (eds.) (2010): Approaches for assessing the impacts of the Rural Development Programmes in the context of multiple intervening factors. European Evaluation Network for Rural Development Working Paper.

Manski, C. F. and Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentenciy and recidivism. *Sociological methodology* 28 (1): 99-137.

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford, New York: Oxford University Press.

Morgan, S. L. and Winship, Ch. (2007). *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Oakes, J.M. (2004). The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social Science & Medicine* 58: 1929–1952.

Pearl J., and Dechter, R. (1996). Identifying Independencies in Causal Graphs with Feedback. In: Kaufmann, M. (ed.), Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference: 420—426.

Pearl, J. (1999). Simpson's Paradox: An Anatomy. Technical Report R-264, Department of Statistics Papers. Los Angeles: Department of Statistics, UCLA. http://bayes.cs.ucla.edu/R264.pdf

Pearl, J. (2000a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2000b). Comment on A.P. Dawid: "Causal inference without counterfactuals". *Journal of the American Statistical Association* 95: 407-450.

Pufahl, A. and Weiss, C. R. (2009). Evaluating the effects of farm programs: Results from propensity score matching. *European Review of Agricultural Economics* 36 (1): 79-101.

Puga, D. (2002). European regional policies in light of recent location theories. *Journal of Economic Geography* 2: 373-406.

Rubin, D. B. (1986). Which Ifs have Causal Answers. *Journal of the American Statistical Association* 81: 961-962.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, Series B, 13: 238-241.

Skoufias, E. (2005). PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico. Research Report 139, International Food Policy Research Institute. Washington, D.C: IFPRI.

Steyvers M., Tenenbaumb, J. B., Wagenmakers, E.-J. and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science* 27: 453–489.

Weaver, W. (1948). Science and Complexity. *American Scientist* 36 (4): 536.