European Commission

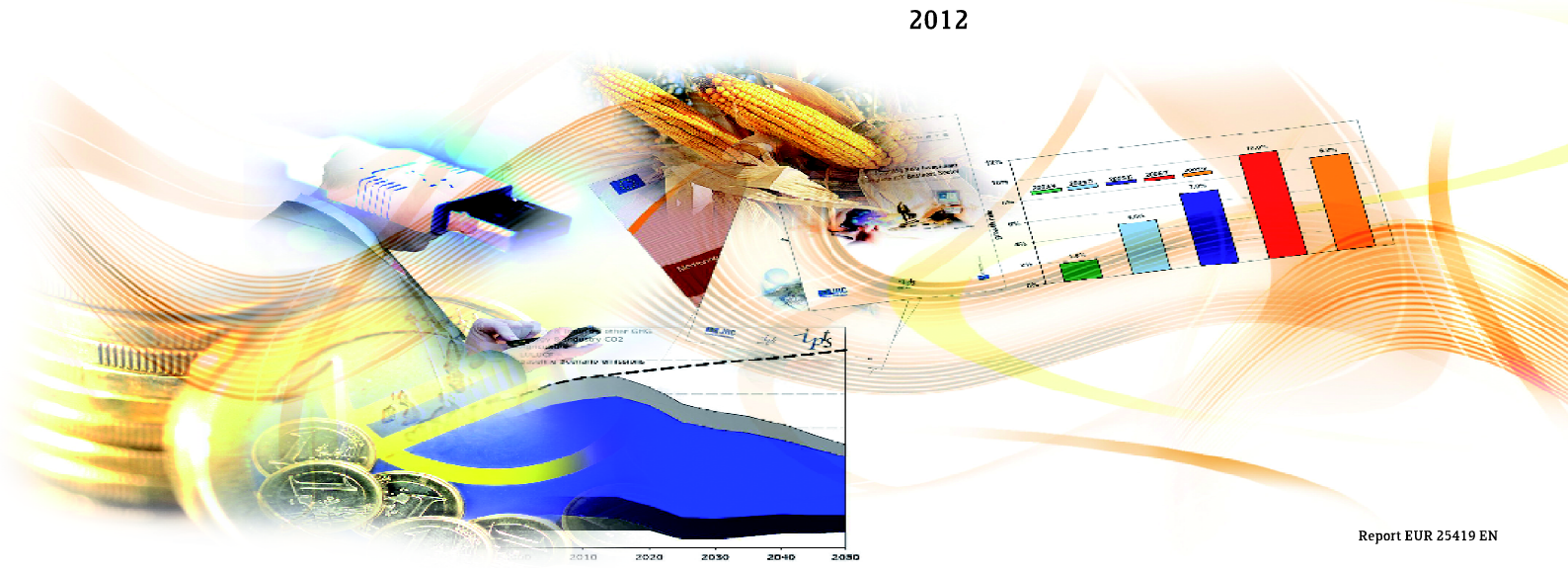
JRC SCIENTIFIC AND POLICY REPORTS

Counterfactual impact evaluation of EU rural development programmes - Propensity Score Matching methodology applied to selected EU Member States

Volume 2: A regional approach

Author
Jerzy Michalek

2012



Report EUR 25419 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for Prospective Technological Studies

Contact information
Address: Edificio Expo. c/ Inca Garcilaso, 3. E-41092 Seville (Spain)
E-mail: jrc-ipts-secretariat@ec.europa.eu
Tel.: +34 954488318
Fax: +34 954488300

http://ipts.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

This publication is a Reference Report by the Joint Research Centre of the European Commission.

JRC 72060

EUR 25419 EN

ISBN 978-92-79-25678-3

ISSN 1831-9424

doi:10.2791/8228

Luxembourg: Publications Office of the European Union, 2012
Rome: Food an Agriculture Organization of the United Nations

Printed in Spain

Counterfactual impact evaluation of EU rural development programmes - Propensity Score Matching methodology applied to selected EU Member States

Volume 2: A regional approach

Author:
Jerzy Michalek[1]

2012

1 European Commission, Joint Research Centre (JRC), Institute for Prospective Technological Studies (IPTS), c/Inca Garcilaso 3, 41092 Seville, Spain. The views, opinions, positions or strategies expressed by the authors and those providing comments are theirs alone, and do not necessarily reflect the views of the European Commission

EUR 25419 EN

Table of Contents

List of Figures

List of Tables

1. EU approach to evaluation of RD programmes

In recent years the evaluation of EU co-founded programmes was assigned particular importance. After the administrative reform of the European Community (Agenda 2000) a periodic evaluation has been extended to all EU policies (Toulemonde et. al., 2002) and recognized as a crucial component of policy development. At the same time evaluation practice became an integral part of EU programming at all levels, e.g. EU, national, and territorial, etc. (Vanhove, 1999; Ederveen, 2003; EC, 1999, 2001, 2002a, 2002b).

According to EU definition, programme evaluation is a process that culminates in a judgment (or assessment) of policy interventions according to their results, impacts and the needs they aim to satisfy[1]. In case of structural and rural development (RD) programmes EU regulations distinguish between ex-ante, mid-term, ex-post and on-going evaluations. Ex-ante evaluations aim at the optimisation of the allocation of the budgetary resources' and the improvement of the quality of programming by answering the question: what impacts can be expected from a newly designed policy intervention?, the main purpose of mid-term and ex-post evaluations of EU programmes is to examine the effects (i.e. results/impacts) of a given programme and to learn about:

- The programme's effectiveness, i.e. the degree to which a program produced the desired outcome (the assessment of a programme's effectiveness implies a pre-definition of operationally defined objectives and their achievement criteria), and
- The programme's efficiency, i.e. the degree to which overall program benefits relate to its costs.

Evaluation literature defines impacts as direct/indirect and intended/unintended effects (economic, social, environmental and others) of a given policy intervention (e.g. development project, programme, policy measure, policy) occurred at *various levels*, i.e. individual, collective or societal and/or local, regional, country, global, etc (i.e. at all possible levels of a "result chain").[2] In contrast, EU evaluation methodology strictly differentiates between programme outputs (physical units), programme results (effects occurred at a micro- level) and programme *impacts*, whereby the last are defined as: *medium/long-term effects of intervention beyond the immediate effects on direct beneficiaries of the programme that can be observed at local community, regional- or macro-economic, country (programme area) or global levels*[3,4]. Following the EU definition, impacts are summative programme outcomes consisting of: a) direct effects on programme beneficiaries (including deadweight loss and leverage effects), and b) indirect programme effects (e.g. substitution, displacement, multiplier, etc.) that occurred at regional, programme area or national levels.

From a policy point of view, impact assessment of a given policy intervention is important as:

1 See: Evaluating EU activities – A practical guide for the Commission Services, DG Budget, July 2004.

2 While analysis of impacts usually distinguishes direct, indirect and induced impacts, definition of impacts differs according to EU terminology vs. World Bank, general evaluation or NONIE terminologies.

3 Rural Development 2007-2013. Handbook on common monitoring and evaluation framework, Guidelines note N. Glossary of terms, EC, 2006

4 In contrast, World Bank, NONIE and other general evaluation guidelines define "impact" broader, by including also direct effects of a given policy intervention at beneficiary level.

- It provides empirical evidence on whether specific policy worked or did not work. It also provides information about the sustainability of effects of a given policy intervention.

- A comparison of a policy intervention's results with target values provides information on the effectiveness of a given policy intervention and on the achievability of more general societal goals (e.g. concerning growth or development) using this specific policy instrument.

- It helps to re-design a policy intervention (programme) to make it more effective and efficient (by taking into consideration costs of intervention).

- It provides arguments for continuation or discontinuation of policies/programmes by comparing social benefits with costs of specific policy interventions.

- It helps to learn about the functioning of economic, social and environmental processes.

- It improves institutional capacities of organisations involved in impact evaluations.

- It improves decision making at all levels.

- It provides some information regarding accountability of institutions involved in the formulation and implementation of policies.

As the evidence for impacts is usually provided on the basis of impact indicators, any appropriate impact assessment should reveal the extent to which observed changes in pre-selected impact indicators (computed at the regional- or macro-levels) came about due to programme activities.

Keeping this in mind, the key challenges of an effective impact assessment of RD programmes carried out at **the regional or macro-levels** are:

Firstly, determining true causation[5], i.e. verifying that an observed change (at micro- or regional levels) of a certain phenomenon (impact indicator) that might be theoretically (!) associated with a given policy (whole or in part) can indeed be attributed (as a whole or partly) to (or is caused by) this policy intervention. In order to verify the above supposition, effects of other intervening factors, (i.e. exogenously determined) which may also influence an observable phenomenon (impact indicator) have to be separated ("netted out") from the effects of this given policy intervention[6]. Such a separation of programme effects from other factors requires a construction of an appropriate counterfactual base-line scenario (a situation without the programme in place).

Secondly, aggregation of various effects of a programme. A summative evaluation of an RD programme's impact should ideally embrace all important programme effects in economic, social, environmental, etc., RD domains and not focus on some programme outcomes only, in form of selected impact indicators (e.g. value added, employment, etc). This can be done by: a) carefully stating the hypothesized effects; b) identifying various possible intended and unintended; direct and indirect; or positive and negative effects that might be caused by a RD programme; c) defining respective measurement criteria; d) defining appropriate time periods to be analysed; and e) systematically monitoring programme implementation. Furthermore, the aggregation of overall programme effects can only be carried out once a consistent weighting system (for individual

5 Causation cannot be proved through a simple correlation analysis.

6 In some evaluations of policy intervention, "impacts" are "identified" as a degree to which certain policy/societal goals (usually pre-defined prior to a policy intervention) have been achieved, after policy intervention. This approach is however not defendable. In fact, certain policy/general societal goals can be achieved without a specific policy intervention via other (policy independent) factors. In this example, an objective of an impact analysis would be inter alia a verification of causality between a degree to which policy goals were achieved (and measured by specific impact indicators) and a given policy intervention

programme effects in various rural development domains) has been developed.

Thirdly, a comprehensive quantitative programme impact evaluation should involve a cost-benefit analysis (including an assessment of the programme's private and social costs and benefits) to be carried out via aggregation and weighting of all partial benefits and costs linked to a given programme.

In order to facilitate evaluations of RD programmes (and ensure a standardized evaluation approach) a common evaluation framework to EU RD programmes was developed by EC (DG-AGRI)[7]. The core element of the EC evaluation framework are Common Evaluation Questions (CEQ) (pre-defined by EC) and programme specific questions (defined by national programme authorities), both to be answered by external programme evaluators. Answering the EC common evaluation questions (CEQ) requires using the concept of "intervention logic," pre-defined by EC, i.e. differentiating between programme inputs, outputs, results, and impacts (by moving from a micro-level to regional- or country levels).

Among dozens of various evaluation questions included in the evaluation guidelines for EU RD programmes implemented in the years 2000-2006 important CEQs concerned an overall effect of implemented policies (e.g. impact on the quality of life)[8]. While impacts of RD programmes at a regional/macro level can occur at various RD domains (economic, social or environmental) programme evaluators were asked to:

- derive their findings using various **partial indicators** describing the potential programme's impact at various RD domains (e.g. economic, environment, etc.) and,
- assess programme net effects by comparing these indicators with respective **common indicators/performance standards.**

The above guidelines have been followed in all evaluation studies of RD programmes implemented during 2000-2006. Yet, many empirical impact evaluations, due to their methodological weaknesses, appeared to be insufficiently rigorous and stringent to serve as a guide to policies.

Clearly, application of inadequate methodologies (e.g. naive methods or absence of control group assessments) for evaluations of programme impacts may lead a number of negative consequences:

- Obtained evaluation results may be heavily biased in both directions (negative or positive). In an extreme situation, results obtained from programme evaluations may substantially differ from real programme impacts (a qualitative difference!).
- Lack of appropriate knowledge about the real impacts of the programme may encourage implementation programmes which, due to their low effectiveness/efficiency, should be discontinued or substantially re-designed.
- Indirect effects of a programme in question may have a decisive impact on the sign of calculated programme net effects. In extreme situations, negative side effects (e.g. economic, environmental, social, etc.) of badly designed RD programmes may impede development of rural areas. Impact methodologies which do not embrace

7 European Commission Agriculture Directorate-General, "Guidelines for the Mid-Term Evaluations of Rural Development Programmes 2000-2006 Supported from the European Agricultural Guidance and Guarantee Fund," 2002; European Commission DG AGRI, "Guidelines for the mid-term evaluation of rural development programmes funded by SAPARD 2000-2006," 2002.; EC, "Rural Development 2007-2013. Handbook on Common Monitoring and Evaluation Framework. Guidance document". Guidelines note N. Glossary of terms, September 2006. http://ec.europa.eu/agriculture/rurdev/eval/guidance/note_n_en.pdf

8 An example of a relevant CEQ can be: "To what extent has a given RD measure/programme contributed to improving of the quality of live in rural areas". The answer to this and other CEQ are to be provided in quantitative terms.

analysis of other indirect effects may lead to inappropriate policy conclusions.

- Poorly designed programmes lead to inefficient allocation of public and private resources and do not contribute to the achievement of policy objectives (e.g. may stimulate sectoral inefficiency, lead to deterioration of competitiveness, and bring about regional divergence). Lack of knowledge about real programme impacts may reinforce those negative developments.

- Insufficient learning about the real programme effects may call into question the credibility of EU evaluations and the institutions involved (conclusions of evaluation reports can be used selectively to support the interest of particular groups or can be contested where the evaluation does not conclude in their favour).

2. The main methodological problems in evaluation studies carried out at macro- and/or regional levels

Numerous ex-post evaluation studies carried out at the regional and macro-levels confirm the existence of huge methodological difficulties faced by evaluators of RD programmes when attempting to:

i) Provide an empirical evidence of a *true cause-and-effect* link between the change in selected impact indicators and the RD programme;

ii) Disentangle for each separate impact indicator (economic, social or environmental) the effect of the RD programme from other exogenously determined factors;

iii) Aggregate and measure the overall effect of an RD programme; and

iv) Perform cost-benefit analysis of the programme.

The major causes of the above difficulties are:

- **Extensive use of traditional evaluation techniques**. Typically, the changes in selected impact indicators (collected at a regional- or macro-level) observed by programme evaluators depend on a number of other (i.e. programme independent) factors (e.g. economy-wide factors, community and household characteristics, social and physical infrastructure activities carried out and supported by other programmes). In this context, calculation of **the net effect** of a given RD programme, i.e. disentangling the effect of a program support from other exogenously determined factors at the regional/macro level definitely cannot be carried out using traditional "naïve" evaluation techniques (e.g. after-before methods). As programme effects cannot be directly observed (see: Chapter 3.1 below) the calculation of a programme impact at a regional- or macro-level requires the application of rigid modern evaluation methodologies and an **obligatory construction of appropriate counterfactuals (i.e. base-line scenario)** (an area which until recently was almost completely ignored by evaluators of RD programmes).

- **Aggregation problems and unclear interpretation in case of opposite or dissimilar effects**. In the majority of cases, effects of a given RD programme in a rural region are multidimensional, i.e. even a single programme measure (e.g. investment in agricultural holdings) can simultaneously affect various RD domains, e.g. production, income, investment, employment, competitiveness, environment, technical and social infrastructure, etc. Additionally, many RD programme measures can have both intended (usually expected by policy makers) and unintended effects**.** For example, investments in rural infrastructure or in processing facilities, along with some positive effects (e.g. increase of labour productivity), may bring about negative environmental impacts, including potential loss of land supporting biodiversity, protected habitats and/or species, deterioration of soil, water environment and air quality, etc. Similarly, support of local food processors may lead to negative effects in the form of strengthening local monopolies (e.g. large processors), causing breakdown of other local food processing businesses, and therefore a decrease of employment and income in non-supported local enterprises, an increase of out-migration, etc.; some investments in irrigation may cause depletion of water

resources in other areas, etc.; support provided to certain type of agricultural producers may have negative effects on on-supported population, etc. In all these cases an assessment of an **overall impact** using pre-selected common impact indicators may be (even for a single RD measure!) rather unmanageable as various effects (positive and negative, expressed in the form of partial indicators) can only with difficulty be compared and/or aggregated (social weights of individual effects in various RD domains e.g. economic, social and environmental are usually unknown). In this context, the partial impact indicators (7 common and 15 additional impact indicators) proposed in the new EC evaluation guidelines for the assessment of an overall net-impact of a RD programme seems to be problematic.

- **Use of average performance standards**. If programme impacts are the main objects of policy concern, reliance on average (regional or country's) performance indicators/ standards as proxy for the functioning of a programme control group can be very problematic. Numerous studies showed that a country's average common performance measures (e.g. average employment rates, growth of income etc.) may not adequately represent a counterfactual situation (i.e. a situation without the programme in place). The evaluation literature suggests that performance standards cannot substitute an econometric impact evaluation based on a comparable control group.

- **Ineffective monitoring system**. The use of various indicators targeting potential effects of specific measures is in practice not possible without having an effective monitoring system (which has to be set up prior to the programme). Yet, the learning about the overall programme effects depends upon *which* (of the possible many) partial indicators are pre-selected and included into the monitoring system. By not including certain indicators, many important impacts (positive/ negative) can be overseen. In order to avoid such situation, a right and timely pre-selection of various partial monitoring indicators and institutional capacity building of monitoring institutions are of crucial importance.

- **Increasing complexity of RD policies**, both in terms of the number of programmes as well as number of applied measures, obviously calls for a multi-dimensionality of evaluation exercise. Given this complexity, estimation of an overall effect of all programme measures (e.g. the effect of the programme support on the quality of life of the beneficiary population) that may simultaneously influence economic, social and environmental domains of rural development requires combination of rigid evaluation methodologies with techniques allowing for a consistent aggregation of impacts by all measures.

Obviously, the **key issue** in evaluation of programme impacts (as well as results) is a **construction of an appropriate counterfactual.** Taking this as a basic criterion, methods used in programme impact evaluations can be divided in four groups (Baker, World Bank 2000; Kapoor, World Bank 2002):

1. Approaches with no counterfactual (e.g. qualitative studies that assess effects of the programme before, during, and after policies are implemented through focus groups, interviews, and other qualitative techniques; "Before and After," methods which compare the performance of key variables during and after a program with those prior to the programme.)

2. Approaches that generate counterfactuals through multiple assumptions (e.g. Computable general equilibrium models (CGEs), regional econometric models, or regional input-output models that attempt to contrast outcomes in treatment and

comparison groups through simulations. While all of these approaches have numerous weaknesses CGE models can produce outcomes for the counterfactual.

3. "Naive" approaches which compare the observed changes in selected performance indicators in a sample of programme areas with arbitrary selected comparison groups.

4. Statistical/econometric methods that control for the differences in initial conditions and policies (both at micro- as well as macro/ regional levels).

Unfortunately, in the majority of studies concerned with the quantitative assessment of socio-economic impacts of RD programmes in EU countries (programming period 2000-2006) "naïve" approaches were employed as a basic evaluation methodology. While in some evaluation studies the authors attempted to build on counterfactuals, in most cases comparisons between supported and non-supported units or areas were done without any consideration for *appropriate* matching. Usually, comparison groups were selected arbitrarily, leading to quantitative results that were statistically biased (i.e. selection bias). In the majority of qualitative evaluations, knowledge about a specific programme's indirect effects (e.g. substitution, displacement, multiplier, etc.) was "imputed" on the basis of anecdotal evidence or ad hoc surveys of a group of beneficiaries, opinions of administrative officials, etc.[9] Furthermore, in approximately 75% of Mid-Term Evaluation (MTE) studies submitted to European Commission by the end of 2010 the impacts of EU RD programmes were assessed *without any reference to a counterfactual situation* (see: EC, European Commission, 2011).

Taking into consideration that in the programming period of 2007-2013 in each individual EU rural region:

- The number of potentially applicable RD measures under an RD programme can be very large (currently up to 42 RD measures can be applied);

- Specific RD measures implemented under specific RD programme will probably affect a wide range of various rural development domains (e.g. economic, environmental, social, etc.); and

- Only seven common partial impact indicators have been proposed to be used for the analysis of impacts of RD programmes (e.g. no common environmental impact indicators are proposed to be used in evaluations of RD measures under Axis 1 and Axis 3; no common economic impact indicators are proposed to be used in evaluations of RD measures under Axis 2);

it is understandable that the assessment of an overall impact of an RD programme at regional or macro-levels requires an application of a more comprehensive and rigid methodological approaches.

9 .See CEAS, 2003. These techniques, in a combination with the most popular "naïve" approach to answering CEQ questions (e.g. the before and after approach) appear as particularly problematic.

3. Advanced empirical approaches

Concerning the use of methodological approaches for impact-analyses of RD/structural programmes *that enable construction of counterfactuals*, the practical possibilities are as follows:

1. The first possibility is to integrate a micro-economic approach (e.g. micro economic individual behaviour or household models) into various local or regional models (e.g. Input-output, Social Accounting Matrix or CGE) and assess the impact of a programme on the base of these combined models (e.g. micro-simulation models with local/regional CGE, village CGE, etc.). The main advantage from the use of these models is a theoretical possibility to estimate both anticipated as well as non-anticipated effects; direct effects (at the beneficiary level) and indirect effects (generated from supply of materials, goods and services attributable to other linked and not directly benefiting units and/or industries located in the same area as well as induced effects (i.e. multiplier effects) of a given programme generated through direct and indirect activities (including consumption, taxes, etc.) of a given policy in question (above models are subject to consistency checks through micro-macro consistency equations). The main disadvantages of these models are: i) input-output models assume that technological/economic relationships are fixed over time and do not respond to price/cost changes; ii) while input-output tables are normally available at relatively high aggregation levels their rescaling to a local level requires a usage of various (often non-transparent) procedures which can be divided in three main categories: "survey", "non-survey" and "hybrid" approaches, e.g. location quotient approach (Del Corpo, et. Al, 2008); iii) commonly applied CGE models usually do not show a detailed enough level of sector disaggregation (a major problem in evaluating RD policies) and are usually static (by contrast, multi-sector and regional dynamic CGE models are much more complex and time consuming in their construction and are therefore very rarely applied to policy evaluations at regional levels); iv) empirical CGE modelling at regional level often is often impossible due to the lack of relevant statistical data at the local or regional level; v) in CGE modelling a heterogeneity of firm behaviour is largely ignored. Despite these deficiencies, micro-macro models are increasingly applied to policy analysis and include a whole array of respective techniques, starting with the simpler macro models that use representative household groups to link macro economic policies and microeconomic data, to more complex top-down modelling frameworks that combine (top) macro models and (down) micro-simulation models (Bourguignon, et al. 2008).

2. The second possibility is to use standard regional input-output econometric models (e.g. REMI, IMPLAN, RIMS II or EMSI) in regional policy analysis to estimate direct, indirect and induced effects of a given policy. For example, the REMI model, that has been in a continuous development since the 1980s integrates input-output, CGE and economic geography methodologies. It consists of thousands of simultaneous equations and its structure consists of five major interrelated blocks: (1) Output, (2) Labor and Capital Demand, (3) Population and Labor Supply, (4) Wages, Prices, and Costs, and (5) Market Shares. The REMI model was applied in numerous studies of economic development in the US and Europe, e.g. for the evaluation of land use and growth controls, impact of investments in energy sectors, transportation,

etc; for the evaluation of regional economic effects of investments in the EU (Treyz F. and G, Treyz, 2002); and recently for an ex-ante evaluation of RDP in Tuscany until 2020 (REMI-IRPET) (Felici, et. al, 2008). The recently extensively used IMPLAN model (the computer software and data-package is available from the Minnesota IMPLAN Group, Inc.) is a computer software package that consists of procedures for estimating local input-output models and associated databases. A Description of the EMSI model is available in: Galloway, H. EMSI's Input-Output Model Multipliers: A Brief Overview and Comparison with Other Major Models, www.economicmodeling.com. Extensive comparison of multipliers used in the REMI, IMPLANT and RIMS II models is available in: Rickman and Schwer, 1995. Yet, the applicability of these models the context of EU policies evaluation raises several concerns. Firstly, it is not quite clear how a number of US economic parameters used in these models can be applied to the EU reality, given different economic and social context in both economies (including problems with data classification and consistency) (comp. Wilson R. in: OECD, 2004); Secondly, modification of these models to reflect local circumstances is usually a considerable and highly time consuming effort that cannot be undertaken by a few external evaluators alone, but requires a great dose of cooperation with local authorities and local stakeholders; Thirdly, the complexity of use for models like REMI or LEFM undoubtedly requires a certain minimum level of expertise; Fourthly, problems with timeliness of the key data incl. input-output tables raises questions regarding forecasting validity.

3. The third possibility to learn about an effect of the programme at the regional- or macro- level is to use a micro- approach and to aggregate direct and indirect impacts computed at the micro-level by drawing on the principles of controlled experimentation (e.g. quasi-experimental approach). This can be done by measuring an individual response (individuals, households, farms, or areas) in controlled settings. Because the supported groups and the comparison groups may differ in observed and unobserved variables that determine programme outcomes, a simple comparison of outcomes between supported and arbitrary selected non-supported units will not reflect the true effect of the programme. To enable such comparisons various techniques can be applied to find adequate controls (e.g. matching; for details see propensity score techniques below). The next step is to derive some meaningful micro-based policy parameters using available data on units in a given sample, e.g. SATE (sample average treatment effect), SATT (sample average treatment on treated), STNT (sample effect on non-treated) and then (by drawing on probability distributions) estimate aggregated impacts for the population at large, e.g. PATE (population average treatment effect), PATT (population average treatment effect on treated), or ATNT (average treatment effect on non-treated), (see: Imbens and Wooldridge, 2007). In many cases, PATE combined with additional information on general equilibrium effects (including substitution and replacement effects) and programme costs (e.g. administrative costs and social costs) can be helpful in answering the policy question regarding the net programme gain to the region, programme area or economy.

4. The fourth possibility is to use an evaluation technique that is based on the matched comparison of regional units (van de Walle, D., and D. Cratty. 2002; Lokshin and Yemtsov, 2005; Michalek, 2008).

Given numerous pros and cons of alternative evaluation methods, it can be particularly advantageous to apply quasi-experimental methods which basically draw on a micro-approach applying it to macro-units (Point 4), i.e.

using a technique that is based on counterfactual analysis involving comparison of **regional** units (van de Walle, D., and D. Cratty. 2002; Lokshin and Yemtsov, 2005; Michalek, 2008). In our study we will follow this approach.

The sequence of analytical steps is as follows:

Firstly, the Rural Development Index (RDI) will be used as the main synthetic impact indicator (Michalek, 2008) - a proxy describing the overall quality of life in individual rural areas. The weights of economic, social and environmental domains entering the RDI are in our study derived empirically from the econometrically estimated intra- and inter-regional migration function after selecting the "best" model from alternative model specifications (i.e. the panel estimate logistic regression nested error structure model, spatial effect models, etc).

Secondly, the impact of RD measures implemented in specific rural regions is analysed by means of selected impact indicators in programme supported regions and control regions, prior to the programme and after it, by applying a combination of the Propensity Score Matching (PSM) (e.g. Kernel matching) and difference-in-differences (DID) methods. Evaluation of programme results at regional levels are performed on the basis of the estimated Average Treatment Effects (ATE), Average Treatment on Treated (ATT) and Average Treatment on Untreated (ATU) effects using the RDI as the main impact indicator.

Thirdly, sensitivity analysis (Rosenbaum bounds) is carried out in order to assess a possible influence of unobservables on obtained results.

Fourthly, given information on regional intensity of programme exposure (financial input flows) the overall impact of the programme support in a selected country is estimated by means of a dose-response function and some derivative dose-response functions under the framework of a generalized propensity score matching (GPS) (Imbens, 2002; Lechner, 2002; Imai and van Dyk, 2002; Hirano and Imbens, 2004). The proposed methodology permits testing a number of common stipulations, e.g. positive effect of a given programme on various indicators of regional performance, e.g. employment, labour productivity, environmental and social indicators, etc.

4. Applied methodologies for evaluation of programme impacts at a regional/macro level

4.1. Fundamental evaluation problem

The main purpose of ex-post evaluation of EU RD programmes is to assess the impact of this policy intervention on regions or programme areas **(i)**, where the programme was implemented.

Similarly as in the case of individuals, the effect of a given EU RD programme on a respective region (or programme area) **i** can be written as:

$$\tau_i = Y_i\,(1) - Y_i\,(0) \qquad (1)$$

Where: τ_i = measures the effect of programme participation on region **i,** relative to effect of non-participation, on the basis of a response variable **Y** (impact indicator). Obviously, as τ_i measures the effect of programme participation for a given region **i**, and **i** is not a subject to any experimental study, only **one** of the potential outcomes, i.e. either Y_i (1) or Y_i (0) **can be empirically observed** for each individual unit/region **i**.

In another words, the fundamental evaluation problem or "fundamental problem of causal inference" arises from the fact that the main policy interest, i.e. the effect of the policy intervention on regions, programme areas, etc. affected by the programme **cannot** be **directly observed** in non-experimental evaluation studies (it is physically impossible to observe the value of the response variable (**Y**) for the same unit/region **i** under two mutually exclusive states of nature, i.e. participation in the programme and non-participation (T*he Fundamental Problem of Causal Inference* (FPCI): Holland, 1986; Rubin 1974; Roy, 1951).

While the FPCI makes *observing* causal effects impossible, this does **not** mean that **causal inference is impossible**. In fact, determining **unobservable** outcome in (eq.1) called *counterfactual outcome is both possible and feasible* (Rubin, 1974; 1975). The literature has long recognized that impact evaluation is essentially a problem of missing data (Ravallion, 2005; Goldstein, 2007) and determining the counterfactual is widely considered the core of each evaluation design (!)[10].

4.2. Policy evaluation indicators

4.2.1. Average Treatment Effects (ATE)

The first indicator which can be applied for evaluation of RD programmes is the **average treatment effect (ATE**). This indicator is simply the difference between the expected outcomes after participation in the RD programme and the outcomes of non-participation conditional on **X** (Heckman, 1996; Imbens, 2003; Imbens and Wooldridge, 2007).

$$\Delta^{ATE}(x) = E(\Delta|X = x) \qquad (2)$$

10 Generally speaking, there are two major methods to determine the counterfactuals, i.e. experimental design and quasi-experimental design. In the experimental design that is generally viewed as the most robust evaluation approach (Burtless, 1995) one would have to create a control group of units which are randomly denied access to a programme. In this random assignment a control group would comprise of firms/units/individuals with identical distribution of observable and unobservable characteristics to those in the supported group. In such an experiment the selection problem would be overcome because participation is randomly determined (see: Bryson, et. al, 2002). Yet, there is a vast literature showing that social experiments (except of in sociology, psychology, etc.) are often too expensive and may require the unethical coercion of subjects unwilling to follow the experimental protocol (see: Winship and Morgan, 1999). As experimental designs (randomization) in case of evaluation of RD programmes would be extremely cumbersome (for ethical and political reasons) a non-random method (quasi-experimental) will be used in this study. The basic idea behind quasi-experimental methods is that they generate comparison groups that are akin to the group of programme participants by using techniques described above.

where:

$\Delta = Y_1 - Y_0$

X = set of observable specific characteristics (covariates) of a given region **i** which are not affected by a given programme.

ATE is the effect of assigning participation randomly to every region **i of type X** (assuming full compliance and ignoring general equilibrium effects) and describes an expected gain from participating in the RD programme for a **randomly selected region i from the joined sub-groups/ regions that participated and those that did not participate** in a given RD programme. This policy indicator averages the effect of the programme over all units in the population, including both programme participants and non-participants. The major disadvantage of these indicators is the fact that ATE includes the effect on regions **j** for which the programme **was never intended/ designed** (it may include impact on regions that may even be programme ineligible).

4.2.2. Average Treatment on Treated (ATT)

The most common policy indicator used for evaluation of programme effects is the **average treatment on the treated effect (ATT),** i.e. in our case showing the average impact of a given RD programme on those regions i where the programme was implemented.

ATT effect can be described as:

$$\Delta ATT\ (x) = E\ (\Delta | X=x,\ D=1) \qquad (3)$$

which is equivalent to:

$$E\ (Y1\text{-}Y0|D=1) = E\ (Y1|\ D=1) - E\ (Y0|D=1) \qquad (3a)$$

ATT can also be defined conditional on P(Z): where P is a probability distribution of observed covariances Z (see: Chapter: 4.3.1.1).

$$\Delta ATT\ (x) = E(\Delta | X=x,\ P(Z)=p,\ D=1) \qquad (3b)$$

As (3a) and (3b) are equivalent, the latter formulation will be applied in our study for calculating effects of a given RD programme.

4.2.3. Average Treatment on the Untreated (ATU)

Information about an eventual extension of a given programme to those that were formerly excluded from the programme can be derived on the basis of an average effect on the untreated (ATU) as defined in (3c).

$$E(Y1\text{-}Y0|D=0) = E\ (Y1|\ D=0) - E\ (Y0|D=0) \qquad (3c)$$

4.3. Construction of control groups

As performance of regions **(i)** supported by a RD programme cannot be directly observed in a "non-support" situation (a given region cannot simultaneously be subject and not be subject to the same programme) economic performance of RD supported regions in a "non-support" situation has to be simulated, using more advanced techniques. Construction of an appropriate base-line should provide us with an answer to the question: "what would have been a given outcome for regions supported by an RD programme if the programme had not been implemented?". By comparing outcomes of the performance of supported regions with a control group of regions in two data points; i.e. at the time of support inititaion and after support, we can straightforwardly answer two questions: Q1). What was the effect of exogenously determined factors[11] on the performance of regions which in reality were supported by the programme?, and Q2). What was the effect of the programme support?

Obviously, in the context of empirical non-experimental studies the counterfactuals cannot be estimated directly, in a manner analogous to the one based on randomization. The underlying

11 All factors which influence performance of supported and non-supported regions and are not considered as RD programme related can be called exogenous.

matching methods seek therefore to mimic conditions similar to experiments, so that the assessment of the RD programme impact world be based on a comparison of outcomes for a group of regions where the RD programme was implemented (D=1) with those drawn from a comparison group of programme non-participants.

One of the difficulties commonly faced during formulation of a relevant base-line is the problem of a perfect comparability (ideally, in case of rural development programmes, the same regions which participated in the programme should also be used for simulation of their performance without the programme). As this is however not feasible, it is important to make comparisons in a manner which guarantees that all basic characteristics of regions in which the RD programme was implemented are as much as possible identical with the characteristics of those regions that did not participate (i.e. the statistical probability of receiving support from RD programmes should be the same for supported and non-supported regions in each comparison group[12]).

4.3.1. Matching

Matching is a method of sampling from a large number of potential controls to produce a control group of modest size in which the distribution of covariates is similar to their distribution in the group of participants. Matching is based on the identifying assumption that conditional on some covariates X, the outcome Y is independent of D.

Application of matching to the consistent evaluation of programme effects makes the following two assumptions crucial:

1. **Unconfoundedness** assumption:
   $Y_0, Y_1 \perp D \mid X$

Where: $\perp$ denotes independence

Unconfoundedness - to yield consistent estimates of the programme impact matching methods assume that the outcome in the counterfactual state is independent of participation, given observable characteristics. This assumption implies, that selection is based solely on observable characteristics and that all variables that influence participation and potential outcomes are observed by the researcher (Caliendo and Kopeinig, 2005).

2. **Overlap assumption**: $0 < Pr(D = 1 \mid X < 1$

The overlap assumption prevents **X** from being a perfect predictor in the sense that it is possible to find a counterpart in the non-participant group for each programme participant and vice versa (Caliendo and Hujer, 2005). If there are regions where the support of **X** does not overlap for the participants and non-participants, matching has to be performed over the common support only (i.e. to avoid a situation of lack of comparable units, one can restrict matching and hence estimation of the effect of programme participation to the region of common support, equivalent to an overlap condition). The overlap condition not only rules out the phenomenon of perfect predictability of **D** given **X** but also ensures that units with the same **X** values have positive probabilities of being both participants and non-participants (see: Caliendo and Kopeinig, 2005; Heckman, LaLonde and Smith, 1999). A weaker version of the overlap assumption implies the possible existence of a non-participant similar to each participant[13].

12 See: Part VI: Application of propensity score

13 Following Heckman, Ichimura and Todd (1998), the importance of overlap assumption can be illustrated on example of a situation where for some values of x we have either p(x) =0 or p(x)=1, i.e. in which one would find some units i with covariates implying that those units either always participate or never participate in the programme. If they always participated there would not have counterparts in the comparison group (non-participants). On the other hand, had they never participated, they would never had counterparts in the group of programme participants.

Conditional on the observables Z, outcomes for the regions which did not participate in a RD programme represent what participating regions would have experienced had they not participated in the RD programme (under the assumption that selection into the RD programme is based entirely on observable characteristics).

Various empirical studies show that traditional matching may be rather difficult if the set of conditioning variables Z is large, due to the "curse of dimensionality" of the conditioning problem[14]. Rosenbaum and Rubin (1983) showed that the dimensionality of the conditioning problem can be reduced by implementing matching methods through the use of so-called balancing scores b(Z), i.e. functions of the relevant observed covariates Z such that the conditional distribution of Z given b(Z) is independent of assignment into treatment. One possible balancing score is the propensity score, i.e. the probability of participating in a programme given observed characteristics Z.

4.3.1.1. Propensity score matching

Propensity score matching (PSM) (Rosenbaum and Rubin, 1983) is used in our study to predict the probability of receiving support on the basis of observed covariates for both supported and non-supported regions. The method balances the observed covariates between the supported group and a control group based on similarity of their predicted probabilities of receiving support, e.g. from the RD programme. The aim of PSM matching is to find a comparison group of regions from a sample of non-supported regions that is closest (in terms of observed characteristics) to the sample of those regions where an RD programme was implemented.

For random variables Y and Z and for discrete variable D, Rosenbaum and Rubin (1983) defined the propensity score as a conditional probability of participating in a programme given pre-programme characteristics Z:

$p(Z) \equiv Pr(D=1|Z) = E(D|Z)$ where Z is a multidimensional vector of pre-programme characteristics.

Rosenbaum and Rubin showed that if the participation in programme is random conditional on Z, it is also random conditional on p(Z):

$$E(D|Y, Pr(D=1|Z)) = E(E(D|Y, Z)|Y, Pr(D=1|Z)) \quad (4a)$$

so that

$$E(D|Y,Z)=E(D|Z)=Pr(D=1|Z) \text{ implies } E(D|Y, Pr(D=1)|Z))=E(D|Pr(D=1|Z)) \quad (4b)$$

Where: Pr (D=1|Z) is a propensity score

In other words when Y0 outcomes are independent from programme participation conditional on Z, they are also independent from participation conditional on the propensity score, Pr (D=1|Z). Conditional independence remains therefore valid if we use the propensity score p(Z) instead of covariates Z or its subset (X).

4.3.2. Matching algorithms

As the probability of observing two units with exactly the same value of the propensity score is in principle zero (since p(Z) is a continuous variable) the estimation of desirable programme effects (see below) requires the use of appropriate matching algorithms which define the measure of proximity in order to define programme non-participants who are acceptably close (e.g. in terms of the propensity score) to any given programme participant.

The most commonly used matching algorithms are: Nearest Neighbour Matching,

14 In case Z is if high dimension it is very difficult to find an appropriate match. For example, with just 20 binary covariates, there are 220 or about a million covariate patterns (Rosenbaum, 2004).

Radius Matching, Stratification Matching and Kernel Matching (Cohran and Rubin, 1973; Dehejia and Wahba, 1999; Heckman, Ichimura and Todd. 1997, 1998; Heckman; Ichimura, Smith and Todd, 1998; Todd, 2006).

4.3.2.1. Nearest neighbour matching

In this matching method the region **j** (non-participant) with the value of Pj that is closest to participanting region Pi is selected as the match.

$$C(P_i) = \min_j |P_i - P_j| \text{ , } j \epsilon I_0 \qquad (5)$$

Where: P is a propensity score

The most prominent variants of nearest matching are i) matching *with replacement*, i.e. the unit, which did not participate in the programme, can be used more than once as a match; and ii) matching *without replacement* where respective programme non-participants can match only once. The biggest disadvantage of the nearest neighbour method is that it can result in bad matches if the closest neighbour (the control unit) is placed far away (in terms of the propensity score) from a supported unit.

4.3.2.2. Caliper matching

This method is to be considered as a variation of nearest neighbour matching. A match for a firm i is selected only if:

$$|P_i - P_j| < \epsilon \text{ , } j \epsilon I_0 \qquad (6)$$

Where ε is pre-specified tolerance

By using caliper matching bad matches can be avoided by imposing a tolerance level on the maximum propensity score distance. The disadvantage of this method is the difficulty to know a priori what tolerance level is reasonable (Smith and Todd, 2005).

4.3.2.3. Kernel matching

Kernel matching is defined as:

$$W(i,j) = \frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)} \qquad (7)$$

Where:

W = weights for i and j
G = a kernel function
an = the bandwidth.

Various kernel functions can be used in applied work, such as the Gaussian, the Epanechnikow, biweight (quartic), triweight or the cosine functions. This non-parametric matching estimator (kernel) is especially interesting as it allows for a match of each programme participant with multiple units in a control group with weights which depend on the distance between the participant observation for which a counterfactual is being constructed and each comparison group observation. In this method weights are inversely proportional to the distance between the propensity scores of participants and controls within the common support level (the further away a comparison unit is from the participant unit, the lower the weight it receives in the computation of the counterfactual outcome). The main advantage of this method is that a lower variance is achieved because more information is used[15]. Another useful property of applying this method is the possibility of using standard bootstrap techniques for estimation of standard errors for matching estimators that generally should not be applied when using nearest neighbour matching (Abadie and Imbens, 2004; Todd, 2006).

15 For systematical analysis of the finite-sample properties of various propensity score matching and weighting estimators through Monte Carlo simulation see: Frölich, 2004b.

4.3.2.4. Local linear weighting function

The local linear weighting function (Heckman, Ichimura and Todd, 1997; Smith and Todd, 2003)) can be defined as:

$$W(i, j) = \frac{G_{ij} \sum_{k \in I_0} G_{ik} (P_k - P_i)^2 - [G_{ij}(P_j - P_i)][\sum_{k \in I_0} G_{ik} (P_k - P_i)]}{\sum_{j \in I_0} G_{ij} \sum_{k \in I_0} G_{ij} (P_k - P_i)^2 - (\sum_{k \in I_0} G_{ik} (P_k - P_i))^2} \quad (8)$$

Where:

W = weights

The difference between kernel matching and local linear matching is that the latter includes in addition to the intercept a linear term in the propensity score of a unit i that participated in the programme. This is an advantage whenever comparison group observations are distributed asymmetrically around the treated observation, e.g. at boundary points, or when there are gaps in the propensity score distribution (Caliendo and Kopeinig, 2005).

Obviously, the specification of a matching algorithm hinges on the two basics factors, i.e. definition of proximity (in the space of the propensity score) and determination of weights (weighting function) (Essama-Nssah, 2006). In some empirical studies 1-to-1 or 1-to-n nearest neighbour with calliper matching methods are used as a standard application. In others, the kernel matching is favoured. Empirical comparison of matching methods suggests that their performance can vary case-by-case thus no one method fits all circumstances and is therefore always preferable (Zhao, 2004; Caliendo and Kopeinig, 2005). Though asymptotically all PSM estimators should yield the same results (Smith, 2000), in small samples the choice of matching algorithm can be important (Heckman, Ichimura and Todd, 1997).

4.3.3. Matching selection criteria

Among many methods allowing to assess the matching quality the most popular approaches are: i) standardized bias (Rosenbaum and Rubin, 1985); ii) t–test (Rosenbaum and Rubin, 1985); iii) joint significance and pseudo R^2 (Sianesi, 2004); or iv) stratification tests (Dehejia and Wahba 1999, 2002). If the quality indicators are not satisfactory, some reasons might be misspecification of the propensity score model (Caliendo and Kopeinig, 2005) or failure of the CIA (Smith and Todd, 2005).

4.4. Difference-in-differences estimator (DID)

DID is a traditional evaluation estimator for cases where the outcome data on programme participants and non-participants is available for both "before" and "after" periods (t' and t, respectively), under assumption that the effect of "unobservables" is time invariant. The DID measures the impact of the RD programme by using the differences between programme participants (D=1) and non-participants (D=0) in the before-after situations (i.e. it compares the before-after change of regions which participated in a programme with before-after change of those control regions which did not participate).

The simplified notation for the DID calculation can be described as follows:

$$DID = \{\Sigma (Yit \mid (D=1) - Yit \mid (D=0)) - \Sigma (Yit' \mid (D=1) - Yit' \mid (D=0))\}/n \quad (9)$$

Where:

(Yit | (D=1) – Yit | (D=0)) is the difference in mean outcomes between the n participants and the m matched comparison units after the access to the RD programme and

(Yit′ | (D=1) – Yit′ | (D=0)) is the difference in mean outcomes between the n participants and m matched comparison units at date 0 (prior to the RD programme).

Yet, the DID method fails if the impact of unobservables is not time-invariant so that a group of programme participants (i.e. regions which participated in a given RD programme) and a control group (regions which did not participate) are on **different development trajectories**. The probability of having different development trajectories increases if already from the beginning of the programme the observed heterogeneity of both groups (and therefore selection bias) is large. While propensity score matching can be applied as a control for the selection bias on **observables** at the beginning of the programme, a **combination** of PSM with DID methods (conditional DID estimator – see 3.5. below) allows for a better controlling of the selection bias in **both** observables and unobservables.

4.5. Combined PSM and Difference-in-differences estimator (conditional DID estimator)

The conditional DID estimator (Heckman, Ichimura and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1998; Smith and Todd, 2005) is highly applicable in case the outcome data on programme participants (i.e. regions which participated in a given RD programme) and non-participants (appropriately constructed control group) is available both "**before**" and "**after**" periods (**t′** and **t**, respectively). In our study, the **PSM-DID** measures the impact of the RD programme by using the differences in selected outcome indicators (ATE, or ATT) between programme participants regions **(D=1)** and **comparable** non-participants regions **(D=0**) in the **before-after** situations.

The conditional PSM-DID estimator can be defined as follows:

$$\textbf{PSM-DID} = \{\Sigma\,(Y_{it} \mid (D=1) - Y_{it} \mid (D=0)) - \Sigma\,(Y_{it'} \mid (D=1) - Y_{it'} \mid (D=0))\}/n \qquad (10)$$

Where:

(Yit | (D=1) – Yit | (D=0)) is the difference in mean outcomes between regions participating in the RD programme and the PSM **matched** control units after implementation of a given RD programme and

(Yit′ | (D=1) – Yit′ | (D=0)) is the difference in mean outcomes between regions participating in the RD programme and PSM **matched** control units at date 0 (prior to the beginning of a given RD programme).

Given ATE, ATT or ATU computed in periods: t and t′ the PSM-DID estimator can be expressed as:

$$\text{PSM-DID} = AT_t - AT_{t'} \qquad (11)$$

Where:

AT = ATE or ATT or ATU

A decisive advantage of the conditional PSM-DID estimator, compared with a standard DID estimator, is that by applying this methodology, initial conditions regarding observable heterogeneity of both groups of regions (programme participants and non-participants) that could influence subsequent changes over time are controlled for.

4.6. Sensitivity analysis

4.6.1. Rosenbaum bounding approach

The unconfoundeness assumption about the treatment assignment merely asserts that all variables that simultaneously affect the participation decision and outcome are observed by the researcher. Yet, if there are unobserved variables that simultaneously affect the participation decision and outcome, a hidden bias might arise to which matching estimators are not robust (Rosenbaum, 2002; Becker and

Caliendo, 2007). The approach proposed by (Rosenbaum, 2002) allows to determine how much hidden bias would need to be present to render plausible the null hypothesis of no effect, or in another words, how strongly an unmeasured variable must influence the selection process in order to undermine the implications of a standard propensity score matching analysis (Caliendo and Kopeinig, 2005).

The Rosenbaum bounding approach does not test the unconfoundedness assumption itself, because this would amount to testing that there are no unobserved variables that influence the selection into the programme; instead it provides evidence on the degree to which any significance results hinge on this untestable assumption (Becker and Caliendo, 2007). An extensive discussion of this sensitivity approach can be found in (Aakvik, 2001; Rosenbaum, 2002; Caliendo and Kopeinig, 2005; Becker and Caliendo, 2007).

Following these studies we define probability of participation as:

$$P_i = P(x_i, u_i) = P(D_i=1|x_i, u_i) = F(\beta x_i + \lambda u_i) \quad (12)$$

Where:

Di = equals 1 if an unit i participates in programme
xi = are the observed characteristics for unit i
ui = the unobserved variable
λ = is the effect of ui on the participation decision

the study is free of hidden bias if λ is zero and participation probability is determined entirely by effects of xi. However, in the presence of hidden bias two matched units (with the same observed covariates x) will have different chances of programme participation. While the odds that both units **i** and **j** will participate are given by $P_i/(1-P_i)$ and $P_j/(1-P_j)$ the odds ratio is equal to $[\exp(\beta x_i + \lambda u_i)]/[\exp(\beta x_j + \lambda u_j)]$ which in case of identical observed covariates (implied by matching) reduces (the vector x cancels out) to $\exp\{\lambda(u_i-u_j)\}$. Rosenbaum, 2002 showed that this implies the following bounds on the odds ratio so that either of the two matched units will participate:

$$\frac{1}{e^{\lambda}} \leq \frac{P_i(1-P_j)}{P_j(1-P_i)} \leq e^{\lambda} \quad (13)$$

If the odds ratio differs, i.e. departs from a value of 1 this can only be due to hidden bias. In this sense **eλ** is a measure of the degree of departure from a study that is free of hidden bias (Rosenbaum, 2002; Caliendo and Kopeinig, 2005; Becker and Caliendo, 2007). Sensitivity analysis means therefore examining the bounds on the odds ratio for programme participation that lie between 1/ eλ and eλ.[16]

Sensitivity analysis, as described above, is applied in our study using formal (Mantel and Haenszel, 1959) test statistics suggested by (Aakvik, 2001) and described in (Becker and Caliendo, 2007). Applications of sensitivity analysis for evaluating social programmes can also be found in (Aakvik, 2001; DiPrete and Gangl, 2004; Caliendo, Hujer and Thomsen, 2005; Watson, 2005).

4.7. Generalized Propensity Score Method

Clearly, propensity score matching described above is especially applicable in situations where an RD programme is implemented selectively (i.e. only in some regions, leaving others unaffected). While this situation (i.e. binary treatment) may be in practice limited to only some specific RD measures (e.g. investment in agricultural holdings, environmental measures, less favoured areas, etc.) the standard praxis is that a given

16 With increasing eλ the bounds move apart reflecting uncertainty in test statistics in the presence of unobserved hidden bias.

RD programme (i.e. in form of aggregated measures) is implemented throughout the whole country, i.e. almost all regions are supported. In case the treatment (i.e. exposure to programme participation) is a continuous variable, the previous setting using a binary propensity score matching has to be extended. Propensity score techniques allowing for multi-valued and continuous treatment effects were proposed by (Imbens, 2002; Lechner, 2002; Imai and van Dyk, 2002; Hirano and Imbens, 2004). Hirano and Imbens (2004) extended the unconfoundedness assumption for binary treatment (Rosenbaum and Rubin, 1983) to multi-valued and continuous treatments and defined the generalized propensity score function (GPS) as the conditional density of the actual treatment given the observed covariates. Empirical applications of a GPS to the evaluation of public policies can be found in (Bia and Mattei, 2007; Kluve et. al, 2007).

Hirano and Imbens, 2004 showed that in combination with the unconfoundedness assumption GPS has a balancing property similar to that of the standard propensity score and thus GPS can be used to eliminate any bias associated with differences in the covariates.

In order to estimate a programme effect at various levels of treatment we will apply the GPS method by following an approach described in Hirano and Imbens, 2004. The approach consists of three main steps:

1. Estimation of the GPS as a conditional density of treatment given the covariates by:

   a. estimation of the parameters of the treatment function (conditional distribution of treatment) using maximum likelihood according to:

$$g(T_i) \mid X_i \sim N\{h(\gamma, X_i), \sigma^2\} \quad (14)$$

   b. assessment of the validity of the assumed normal distribution model by appropriate tests (e.g. Kolomogorov-Smirnov, Shapiro-Francia, Shapiro-Wilk or skewness and kurtosis tests for normality)

   c. estimation of the GPS as:

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\hat{\sigma}^2}\{g(T_i) - h(\hat{\gamma}, X_i)\}\right] \quad (15)$$

   where $\hat{\gamma}$ and $\hat{\sigma}^2$ are the estimated parameters in step a).

   d. testing the balancing property

2. Modelling the conditional expectation of the programme outcome as a flexible function (polynomial approximation) of Ti and Ri

3. Estimation the average potential outcome for each level of treatment and an entire dose-response function as:

$$E\{\widehat{Y(t)}\} = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}\{t, \hat{r}(t, X_i)\} = \frac{1}{N}\sum_{i=1}^{N}\varphi^{-1}[\hat{\psi}\{t, \hat{r}(t, X_i); \hat{\alpha}\}] \quad (16)$$

where $\hat{\alpha}$ is the vector of the estimated parameters in the second stage.

5. Impact Indicators

5.1. Rural Development Index

The main summative impact indicator used in this study of evaluation of RD programmes is the RDI. The methodology applied to construction of a synthetic index of the rural development (RDI) is described in Michalek and Zarnekow, 2011; and Michalek and Zarnekow, 2012. The RDI, as a composite indicator, was calculated according to eq. (17) on the base of regional characteristics $\mathbf{Z_i}$ and individual weights $\boldsymbol{\beta_k}$ that were derived from the estimated migration function (see: eq. 18). In such a model, the estimated weights $\boldsymbol{\beta_k}$ represent the relative "importance" or a "social value" assigned by a society (composed of those who migrated and those who stayed) to each of characteristics $\mathbf{Z^k_i}$ representing various aspects of the quality of life in all origin and destination regions **i**.

Formally the RDI in each individual region **i** can be expressed as a linear function of specific **i**-region characteristics $\mathbf{Z^k_i}$ and their weights $\boldsymbol{\beta_k}$ (see eq 17):

$$\mathbf{RDI_i} = h\,(\boldsymbol{\beta}_k, Z_k^i) = \Sigma_k\, \boldsymbol{\beta}_k * Z_k^i \qquad (17)$$

Where:

$\mathbf{RDI}_i$ = Rural development index (an equivalent of the quality of life index) in region **i**

$\mathbf{Z_k^i}$ = Measurable characteristics **k** in a region **i**

$\boldsymbol{\beta}_k$ = Weights for each characteristic **k** derived from the estimated migration function that can be both **i**-region and time **t** specific

In our study $\mathbf{Z_k^i}$ is constructed empirically using factorization method applied to all relevant coefficients and variables $\mathbf{V^i}$ available at the regional level. The latter are nested in $\mathbf{Z_k^i}$ (i.e. RD domains) and describe in detail various specific aspects of rural development in each individual region **i** (e.g. a number of enterprises, employment coefficients, water/air pollution coefficients, schools, health facilities, etc. available from regional secondary statistics).

Weights $\boldsymbol{\beta_k}$ that enter the RDI are derived from a migration model (eq 18) where the probability distribution of migration log (m) is a dependent variable, and differences in regional characteristics $\Delta F_{IDKt'}$, and transaction costs (D) are explanatory variables. While weights $\boldsymbol{\beta_k}$ used to construct the RDI are only a *subset* of estimated coefficients within a migration model, this feature brings about a separation of the RDI from migration (due to transaction costs).

The migration model applied for derivation of weights in the RDI was estimated as a panel regression in form of (18):

$$\log(m)_{ID,t} = \alpha_0 + D_{ID}\cdot\delta_1 + D_{ID}^2\cdot\delta_2 + \Delta F_{IDKt}\cdot\beta_K + v_{ID} + \varepsilon_{IDt}$$

(18)

Where:

$$\log(m) = \log\left(\frac{\text{mrate}}{\text{1-mrate}}\right)$$

mrate = inflows from region i to j divided by (population in i multiplied by population in j)

D_{ID} = distance between region i and j

D^2_{ID} = squared distance between i and j

ΔF_{IDKt} = differences in factors k between regions i j

v_{ID} = random intercept at the pair wise ID level

ϵ_{IDt} = residual with `` usual´´ properties (mean zero, uncorrelated with itself, uncorrelated with D and F, uncorrelated with v and homoscedastic).

$\epsilon = N(0, \sigma^1_\epsilon)$

As a random effect model it assumes the random effects occur at the level of the pairwise migration flows between all regions **ij** (region as a group variable). Model 18 is thus estimated as a random effect linear regression model with a group variable at the level of **i j** (ID) by using the GLS random effects estimator (a matrix-weighted average of the between and within estimators)[17].

The most important pros and cons of selecting Model 18 as a base for derivation of weights used in calculation of the RDI in comparison with other alternative model specifications are provided in Michalek and Zarnekow, 2009.

The major advantages from applying the RDI as an impact indicator to the evaluation of RD programmes are as follows:

- The approach allows to consider all potential effects of a given RD programme (aggregated or separated by programme measures) on various rural development domains (economic, social, environmental, etc.) and on the overall quality of life of population living in individual rural areas.
- The approach allows to incorporate numerous general equilibrium effects of a programme, e.g. multiplier effects, substitution effects, into the analysis .
- As an impact indicator the RDI is powerful both at the aggregated level (e.g. NUTS 2) and commune levels (NUTS 5) and even the village level (if data exists).
- As an impact indicator the RDI is applicable both for analysis of RD programmes as well as analysis of structural programmes.
- The RDI can also be used as an impact indicator for the evaluation of large projects implemented at low regional levels (e.g. NUTS 5).

5.2. Other partial impact indicators

Beyond the RDI, other selected partial performance indicators available at regional level (e.g. employment coefficient, rate of rural unemployment, value added, etc.) were used as relevant impact indicators.

17 The random effect estimator produces more efficient results than between estimator, albeit with unknown small sample properties. The between estimator is less efficient because it discards the over time information in data in favour of simple means; the random-effects estimator uses both the within and the between information (STATA, ver.10; Kennedy, 2003).

6. Synthesis of the methodological approach to the evaluation of the impact of RD programmes

The evaluation techniques described above were applied to the assessment of the impact of an RD programme (SAPARD) in Slovakia and Poland. The following steps were carried out:

Firstly, the RDI (as described above) was computed for all i-regions (i.e. where the RD programme was implemented and non-implemented) in a given country.

Secondly, binary propensity score matching was applied to estimate the impact of individual SAPARD measures (in both countries individual SAPARD measures were implemented in some regions only (programme participants) and not throughout the whole country) using the RDI, and the unemployment rate as impact indicators. Propensity scores for individual regions in a given country were obtained from a standard logit-model with region- and time-specific characteristics (factors/principal components) computed prior to the beginning of the SAPARD programme (2002) as explanatory variables.

Thirdly, some of regions were excluded from further comparisons because their propensity scores were outside the common support. Matched pairs of similar regions etc. were constructed on the basis of how close the estimated scores were across the two samples (supported vs. controls). Several weighting techniques (matching algorithms) were applied to calculate the average outcome indicator of the matched non-supported group, ranging from "nearest neighbour" weights to non-parametric weights (e.g. kernel functions of the differences in scores). The "best" matching algorithm was selected using a minimum standardized bias as a main criterion (conditional on meeting other criteria, e.g. t-tests, ands pseudo R^2 test).

Fourthly, the mean value of the outcome indicator (i.e. RDI and other relevant partial outcome indicators, e.g. unemployment) for the nearest "neighbours" of the programme supported regions was computed using a selected matching algorithm (e.g. Kernel method).

Fifthly, the conditional DID method (combination of PSM and DID) was applied to measure the impact of the RD programme on individual regions (2002-2005)[18].

Sixthly, a sensitivity analysis was carried out in order to find out: i) whether unobserved factors at the regional level could alter inference on effects of participation in SAPARD, and ii) how strongly an unmeasured variable would have to influence the selection process to undermine the implications of the matching analysis. An assessment of a possible influence of unobservable characteristics on procured results was obtained by applying the methodology described in (Rosenbaum, 2002). The approach "Rosenbaum bounds" allows for testing the presence of unobserved heterogeneity (hidden bias) between supported and non-supported regions. The testing procedure is carried out on the basis of Mantel-Haenszel test statistics that give bound estimates of significance levels at given levels of hidden bias under the assumption of either systematic over- or underestimation of treatment effects. The sensitivity analysis was carried out using a syntax described in: Becker and Caliendo, 2007.

18 Specifically, the difference "one" is the difference in mean outcomes between those regions where programme was implemented and the matched comparison regions after implementation of the RD programme, the difference "two" is the difference in mean outcomes between those regions where programme was implemented and matched comparison regions prior to the RD programme, and the difference "three" is the difference between difference "one" and difference "two".

Seventhly, respective impact assessments using above impact indicators were carried out for each individual SAPARD measure separately (a specific base-line was derived for each RD measure).

Eighthly, the net-impact of the whole RD programme (all measures-together) was estimated at various intensity levels of programme exposure (level of programme expenditures) using the RDI as a synthetic impact indicator and the local unemployment rate as an important partial indicator at regional level. A generalized propensity score methodology that allows for continuous treatment regimes was applied to derive the dose-response function and the derivative of the dose-response function.

The above methodology was empirically applied for an estimation of the impact of SAPARD in Poland and Slovakia at the NUTS-4 level in the years 2002-2005.

7. Data:

Poland: Data used for calculation of the RDI at (NUTS-4) in Poland originates from the Regional Data Bank (RDB) of the Polish Statistical Office of the Ministry of Finance (e.g. distribution of personal income) and the Ministry of Interior (e.g. crimes). Above data was collected either at the NUTS-5 level and then aggregated to NUTS-4 or directly at NUTS-4 levels for the years 2002 to 2005. Of 379 NUTS-4 regions in Poland 314 rural Powiats (NUTS-4) are included in the analysis (84.2% of all NUTS4-regions), which excludes 65 big cities. Data basis for Poland covers all relevant rural development dimensions available in regional statistics at the NUTS-4 level and consists of 991 coefficients/indicators collected/calculated either directly at the NUTS-4 level or aggregated from NUTS-5 (approximately 2500 Polish gminas) levels into the NUTS-4 level. Furthermore, above data was supplemented with information on allocation of SAPARD funds (by measures) among NUTS-4 regions. The data base covers the period of 2002-2005.

Slovakia: The database for Slovakia originates from Slovak Statistical Office whereby 337 indicators/variables collected at 72 regions (NUTS-4) are used for construction of the RDI. Furthermore, similar as in Poland, above data was supplemented with information provided by RIAFE on allocation of SAPARD funds (by measures) among NUTS-4 regions. The data base covers the period of 2002-2005.

8. Results

An econometric estimation of weights in the RDI was carried out separately in both countries on the basis of eq. 18. A detailed description of an approach used for the derivation of the RDI in both countries and results obtained can be found in Michalek and Zarnekow, 2011; Michalek and Zarnekow, 2012.

Poland

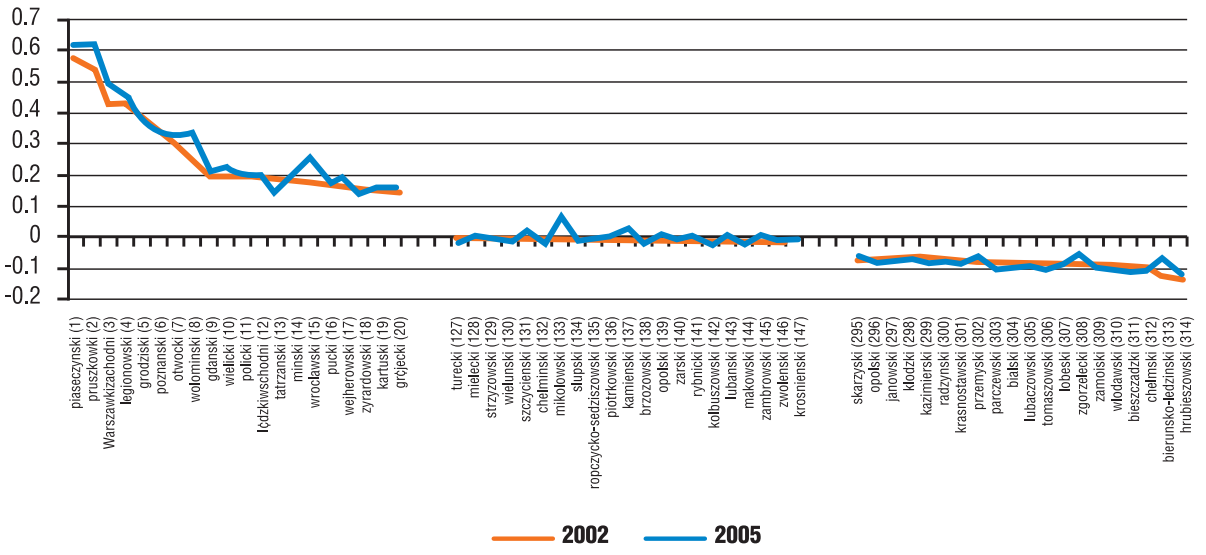
8.1. Construction of the RDI as a programme impact indicator

The RDI in Poland involving 991 regional indicators was calculated for all (314) rural NUTS-4 regions and the years 2002-2005 according to eq. 17. On the basis of the estimated RDIs rural regions were ranked in terms of their overall level of rural development. The ranking of NUTS-4 regions over the years 2002-2005 is shown in Figure 1. The geographical distribution of the RDI in Poland (the average of 2002 and 2005) is shown in Figure 2.

The results of the RDI estimation confirm a clear typological division of Poland based on the performance of individual rural regions into a good performing western- and central part, and a badly performing eastern part (north-eastern and south-eastern),. The results also back up a general opinion that suburbs of the biggest cities (e.g. Warsaw, Poznan, Gdansk, Wroclaw, Lodz, Krakow) exhibit the highest quality of life (see Figure 2). The lowest RDIs (i.e. less than -0.08) were found in remote regions situated in south-eastern Poland, i.e. hrubieszowski (on the border with Ukraine), bierunsko-ledzinski (a former heavy industrial complex in south Poland), chelmski (on the border with Ukraine), bieszczadzki (a remote region bordering to Ukraine and Slovakia) for details see Table 1 in Annex).

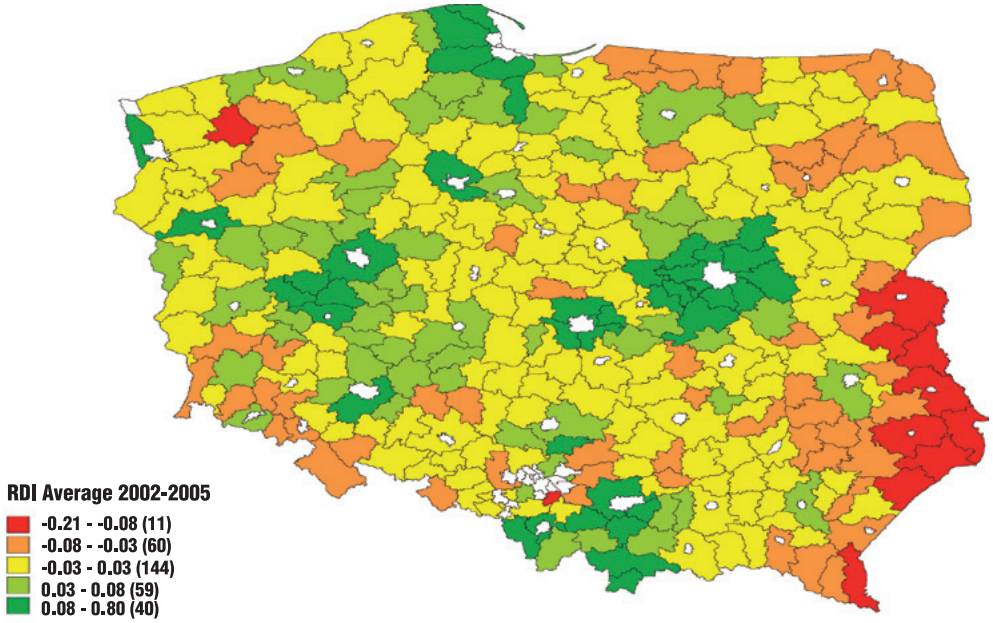
As mentioned before, an estimation of the RDI (by region) in Poland was carried out on the basis of factors obtained by applying a principal

Figure 1: Poland: Ranking of regions. RDI by regions (NUTS-4, 314 regions)



Source: Michalek and Zarnekow, 2009.

Figure 2: Poland: Average RDI (by regions and years 2002-2005)



Source: Michalek and Zarnekow, 2009.

Table 1: Poland: List of individual rural development components (2002-2005)

Factors	Rural development component
F1	Employment by sectors
F2	Lowest income groups and structure of own budgetary resources
F3	Population density and urbanisation
F4	Highest income groups and housing availability
F5	Subsidies and social expenditures
F6	Population structure
F7	Industrialization, investments and fixed assets
F8	Gas supply system
F9	Tourist sector, newly registered companies
F10	Employment conditions and work hazard
F11	Heating energy sector <pollution> and deaths
F12	Natural population growth
F13	Public administration and social infrastructure
F14	Unemployment structure and dwelling equipment
F15	Social sector and its financing
F16	Structure of local budgets
F17	Environmental pollution and infrastructure

component method to 991 regional coefficients showing various aspects of rural development. The same factors (f1-f17) representing individual regional characteristics in the years 2002-2005, are used later as the main covariates explaining differences in regional performance and the probability of the selection of individual regions into specific rural development programmes.

The overview of the main factors/components is shown in Table 1.

Due to its comprehensiveness, the RDI can be used as the impact indicator measuring the effects of *various* rural and structural programmes affecting rural areas. In our study, the RDI will be applied to evaluation of the overall impact of the pre-accession SAPARD programme (2002-2004).

8.2. Scope and regional distribution of the selected SAPARD measure

The assessment of the impact of the SAPARD programme in Poland was carried out by taking as an example a measure that was especially designed to improve the quality of life of the population living in rural areas (i.e. SAPARD measure 3 "Development and improvement of rural infrastructure"). Of 6230 investment proposals submitted under this measure to the Polish Agency for Modernisation and Restructuring (SAPARD implementing agency) 4492 contracts (years 2002-2004) were signed and implemented in the following years amounting to approximately 2 bn PLN (approximately €547m), of which 1.520m PLN (€411m) were co-financed from the EU. The main beneficiaries of this measure were local administration units (gminas at NUTS-5 and poviats at NUTS-4 levels). The major financial allocations under Measure 3 concerned the development and modernisation of roads (41%), waste water disposal (41%), water supply to agricultural holdings (16%), solid waste management (0.41%), and the provision of renewable energy (0.35%).

An impact assessment of a given RD programme (measure 3) requires some basic information about:

1. Which regions were supported by the given RD programme (measure 3)?; and

2. What was the local/regional intensity of this support?

Although basic data on financial aspects linked to the implementation of Measure 3 under the SAPARD programme was generally available (e.g. total programme spending by measure and region) answering the above questions could create some problems because:

a. In several regions (average NUTS-4 region, 81 000 population and 973 km²), funds from the SAPARD programme that were allocated during the years 2002-2004 to eligible infrastructural investments under measure 3 were *almost* negligible (e.g. total public support from this programme measure was less than €0.1m per region). In this situation, it would not be justifiable to classify these regions as *supported* from the programme;

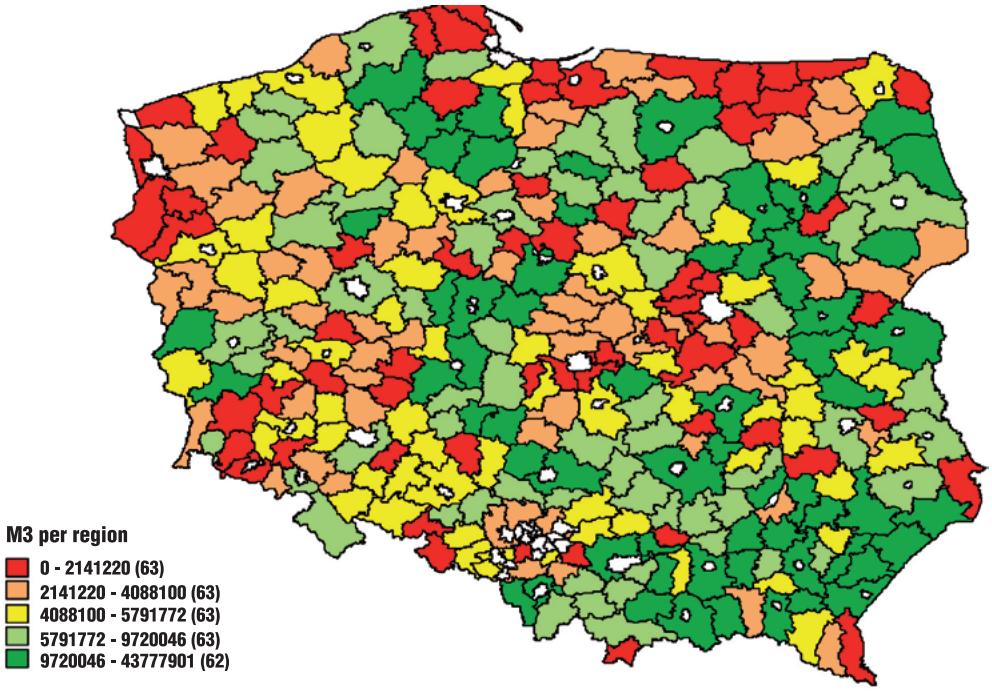
b. The intensity of the programme support can be measured using various indicators, e.g. total per region; per capita in region; or per km² in region. While all of these indicators have both advantages and disadvantages, an objective appraisal of programme impact may require using of all three criteria.

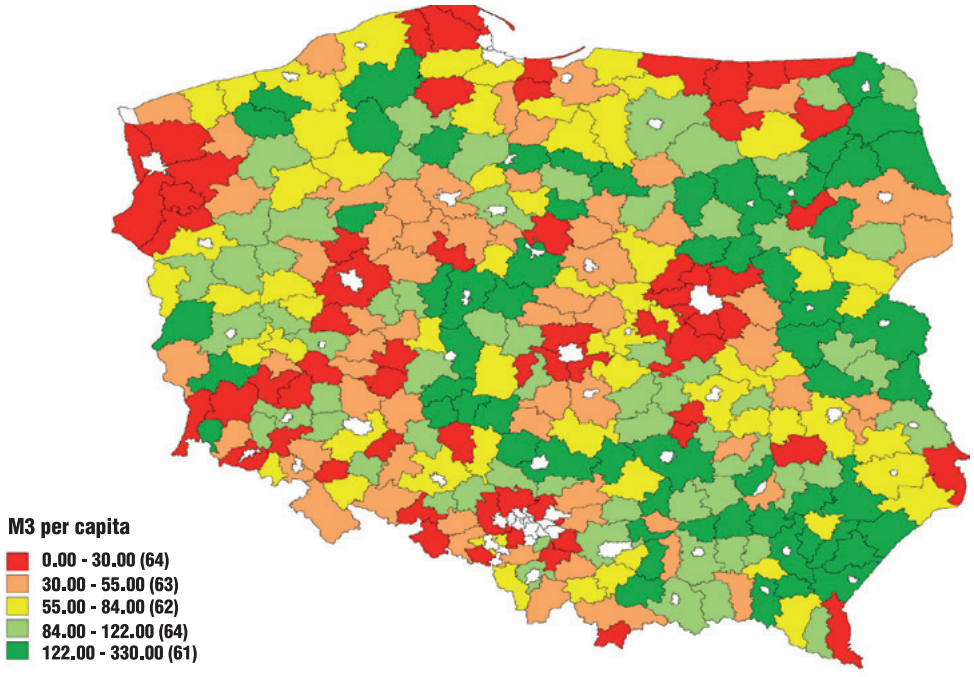
An analysis of the geographical allocation of funds under SAPARD Measure 3 shows that programme resources were not equally distributed across all NUTS-4 regions.

The majority of available resources under SAPARD (Measure 3) were used to improve the rural infrastructure in eastern and south-eastern Poland (see graphs 2a-2c). These were also the areas where individual exposure/intensity (per region, per capita or km²) to the programme (measure 3) was the highest.

Further analysis of allocation of funds under measure 3 shows a negative correlation of the programme intensity with the RDI (see Table 2) thus confirming that available resources from SAPARD (Measure 3) were primarily targeting less- and medium-developed rural regions.

Figure 3: Poland: Allocation of SAPARD funds (Measure 3) by regions





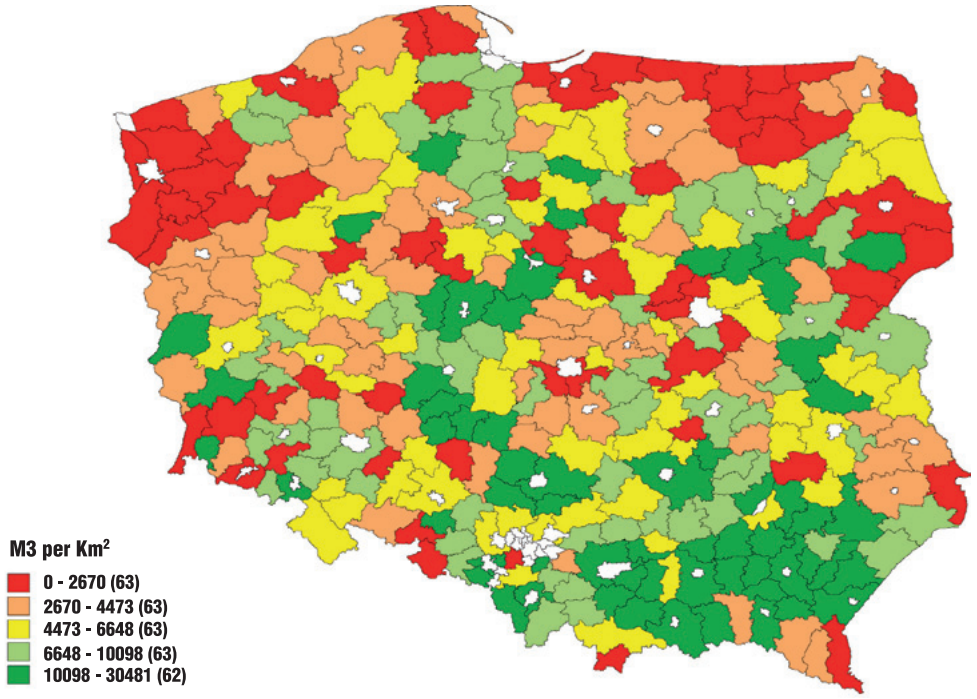


Table 2: Pearson correlation matrix between RDI Index and M3 funds

	RDI 2002	M3	M3_pp	M3_km
RDI 2002	1.0000			
M3	-0.0592	1.0000		
M3_pp	-0.1974	0.7695	1.0000	
M3_km	-0.0128	0.7434	0.7156	1.0000

8.3. Application of the binary PSM matching

8.3.1. Division of regions between supported and non-supported

Considering that only five (out of 314) NUTS-4 regions did not receive any support from the SAPARD programme under Measure 3, and in a further 16 regions the support from the programme (Measure 3) was almost negligible (i.e. did not exceed €200,000 per region), an arbitrary threshold had to be imposed to differentiate between programme supported and non-supported regions. As a general rule, those regions where the programme intensity (Measure 3) was lower than 2/3 of the median were qualified as "not supported". The same rule (removed in Chapter 8.9) was applied to all programme intensity measures (i.e. M3 per region; M3 per capita; and M3 per km²)[19].

19 While, the effectiveness of relatively small yet well designed investments (e.g. addressing point source environmental pollution) can be very high thus setting of a threshold above which amount a region can be considered as supported is always arbitrary.

8.3.2. Intensity of programme exposure per region basis (M3 per region)

The application of an indicator "M3 per region" (and the setting of the above threshold) as the main criterion determining the status of an individual region (supported vs. non-supported) resulted in a division of 314 NUTS-4 regions into two groups: i) programme participants (185 regions), if programme funds (measure 3) allocated to a respective region were above 4.1m PLN (€1.1m) per region; and ii) programme "non-participants" (129 regions), if allocated funds per region were below this threshold[20].

Initial differences in regional characteristics of participants vs. non-participants prior to the implementation of SAPARD (2002) are shown in Table 3.

Table 3: Initial differences in regional characteristics of participants vs. non-participants prior to implementation of SAPARD (2002)

Variable	Mean		Difference (1-0)
	D=1 (185)	D=0 (129)	
f1	-.1408335	.3287903	-.46962387
f2	-1.06515	-1.073821	.00867066
f3	-.0208805	.1059945	-.12687502
f4	-.1198797	.0644299	-.18430958
f5	-.1521219	.1960632	-.34818508
f6	.3772656	.3957994	-.01853381
f7	-.0447577	.1389971	-.18375473
f8	.1659376	-.2083546	.37429221
f9	-.0655768	.1143073	-.17988411
f10	-.0989065	-.0091456	-.0897609
f11	-.1247719	.1451686	-.26994055
f12	.0522556	.1576558	-.10540019
f13	.3525906	-.1479066	.50049721
f14	-.9706903	-1.223194	.25250344
f15	-.0753778	.0332847	-.10866255
f16	.0343757	-.1023875	.13676314
f17	.3453493	.1128787	.2324706
RDI2002	.0132698	.0307774	-.01750756
unemplrur02	.6249013	.5212309	.10367045

20 We note that the use of the intensity of programme exposure per region as the main programme participation criterion may lead to discrimination of small rural regions. Indeed, in extreme situation some small regions (programme participants) could be assigned a status of "non-participants" only due to the fact that allocated programme funds did not exceed the arbitrary threshold (as at region basis). In fact, by setting a threshold the interpretation of the programme impact may change by restricting it to effects of substantial programme allocations (above 1.1 Mill EUR).

Where:

F = endowments in factors/ RD components
D = 1 (Group 1; i.e. programme participants)
D = 0 (Group 2; i.e. programme non-participants
RDI2002 = RDI in 2002
Unemplrur02 = rural unemployment rate (% of rural unemployment in total unemployment)

We note that both groups of regions (Group1 = supported vs. Group2 = non-supported) differed considerably both in terms of their overall level of rural development (measured by the RDI) as well as in terms of other regional characteristics (factors 1-17, total unemployment, rural unemployment, etc.). For example, the overall level of rural development (measured in terms of the RDI prior to the SAPARD programme in year 2002) in the group of regions qualified here as programme participants (i.e. Group 1: less developed regions) was about half of the group of programme non-participants (i.e. Group 2: better developed regions). When compared with the level of rural unemployment (the percentage of rural unemployed in the total unemployed), the respective figures prior to the SAPARD programme were 62% in Group 1 compared with 52% in group 2. The analysis of individual factors characterizing other aspects of rural development prior to beginning of the SAPARD programme, e.g. f4 (percentage of the highest income groups and housing availability), f5 (subsidies and social expenditures), f8 (rural infrastructure, e.g. gas supply system) or f11 (Heating energy sector <pollution> and deaths) indicate significant differences between both groups of regions (see Table 3). It also indicates a much worse economic, social and environmental performance of Group 1 (later supported by SAPARD) compared with Group 2 (non-supported). Given the above, we therefore conclude that the allocation of SAPARD funds (Measure 3) was **carefully targeted** and determined by the actual economic, social and environmental situation of individual rural regions.

Clearly, significant differences in individual characteristics (factor endowments) in both groups of regions prior to the SAPARD programme (2002) confirm the existence of a considerable **selection bias** and therefore a **non-direct comparability of both groups of regions**. In other words, a direct use of selected impact indicators (e.g. an RDI, added value, employment etc.) for assessment of the impact of SAPARD by performing a counterfactual analysis confined to a simple comparison of performance of these indicators in the above groups (e.g. using a traditional DID method) would not be appropriate. This could lead to biased results unless there is strong additional evidence that the hidden bias (unobserved by evaluators) remains time invariant. As this cannot normally be guaranteed, one should apply evaluation techniques that ensure the full comparability of programme participants and control groups of regions, e.g. by drawing on matching principles (e.g. propensity score matching).

8.4. Estimation of propensity score

Given information about individual regional characteristics prior to the SAPARD programme (year 2002) and the status of each individual region (programme participants vs. programme non-participants), a logit function was estimated using factors (f1-f17) and unemployment coefficients as covariates. The results of the logit estimation are shown in Table 4

The results of this estimation were then used to derive the individual probability of programme participation (propensity scores) for all regions. Clearly, in order to ensure comparability, the estimated propensity scores of regions that participated in the SAPARD programme (measure 3) and their controls should be very similar. As the probability of observing two units with *exactly* the same value of the propensity score is in principle zero (since p(Z) is a continuous variable), the estimation of desirable programme effects (e.g. ATT, ATE, etc.) requires using appropriate matching algorithms. These set up the measure of proximity in order to define

Table 4: Poland: Logit estimates (results)

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Unemploy 2002	20.93901	8.441632	2.48	0.013	4.393719	37.48431
f1	-1.035982	.2015297	-5.14	0.000	-1.430973	-.6409913
f2	-.3558105	.9428628	-0.38	0.706	-2.203788	1.492167
f3	-.0246205	.168223	-0.15	0.884	-.3543314	.3050905
f4	.0573888	.2154196	0.27	0.790	-.364826	.4796035
f5	-.4712841	.1674573	-2.81	0.005	-.7994943	-.1430738
f6	.0897225	1.363651	0.07	0.948	-2.582985	2.76243
f7	-.1889602	.1547246	-1.22	0.222	-.4922149	.1142944
f8	.7496832	.1929452	3.89	0.000	.3715176	1.127849
f9	-.2802929	.1503198	-1.86	0.062	-.5749142	.0143285
f10	.0020269	.1481626	0.01	0.989	-.2883664	.2924203
f11	-.5908087	.1793082	-3.29	0.001	-.9422464	-.2393711
f12	-.3111301	.1585878	-1.96	0.050	-.6219564	-.0003038
f13	.6414907	.1562156	4.11	0.000	.3353138	.9476677
f14	.7415563	.2765786	2.68	0.007	.1994722	1.28364
f15	-.1539574	.1511791	-1.02	0.308	-.4502629	.1423482
f16	.2331552	.1500473	1.55	0.120	-.0609321	.5272426
f17	.2636456	.148621	1.77	0.076	-.0276461	.5549373
_cons	-1.159306	1.613543	-0.72	0.472	-4.321791	2.003179

Logistic regression	Number of obs	LR chi2(18)	Prob > chi2	Log likelihood	Pseudo R2
	= 314	= 102.76	= 0.0000	= -161.2499	= 0.2416

programme non-participants who are acceptably close (e.g. in terms of the propensity score) to any given programme participant.

8.4.1. Selection of a matching algorithm

The most commonly used matching algorithms involving propensity score are: Nearest Neighbour Matching, Radius Matching, Stratification Matching and Kernel Matching (Cohran and Rubin, 1973; Dehejia and Wahba, 1999; Heckman, Ichimura and Todd. 1997, 1998; Heckman; Ichimura, Smith and Todd, 1998). While asymptotically all PSM matching techniques should yield the same results, the choice of matching method (or applied matching parameters e.g. number of nearest neighbours, radius magnitude, kernel type, etc.) can make a difference in small samples (Smith, 2000)[21]. As the quality of a given matching technique depends strongly on a dataset, the selection of a relevant matching technique in our study was carried out using three independent criteria: i) standardized bias (Rosenbaum and Rubin, 1985); ii) t–test (Rosenbaum and Rubin, 1985); and iii) joint significance and pseudo R^2 (Sianesi, 2004).

We found that the best results were achieved by using an iterative procedure (e.g. linear

21 Description of trade-offs linked to each of matching algorithms can be found in (Caliendo and Kopeinig, 2005).

Table 5: Poland: Comparison of matching algorithms (participation criterion: M3 per region; impact indicator: RDI in 2002)

Matching method	Matching parameters	Estimated standardized bias (after matching)
Nearest neighbours	N (6)	9.59
	N (7)	8.88 → min
	N (8)	9.73
Radius caliper	(0.2)	7.57
	(0.21)	7.41 → Selection Min {Min}
	(0.22)	7.47
Kernel normal (Gaussian)	bandwidth (0.08)	7.64
	bandwidth (0.09)	7.48 → min
	bandwidth (0.10)	7.57
Kernel biweight		7.92
Kernel epanechnikov	bandwidth (0.25)	7.59
	bandwidth (0.24)	7.58 → min
	bandwidth (0.23)	7.61

search) with a minimization of the calculated standardized bias[22] (after matching) as an objective function and applying min{min} as the main selection criterion. In all considered cases (various matching algorithms)[23] an optimal solution could easily be found due to local/global convexity of the objective function with respect to function parameters under each matching algorithm (e.g. radius magnitude in radius matching; or number of nearest neighbours in nearest neighbour matching). An overview of results obtained using different matching algorithms is provided in Table 5.

In our example (314 total observations; participation criterion: M3 per region; impact indicator: RDI in 2002) the radius calliper matching (0.21) was selected as the best matching algorithm (see Table. 5). The imposition of a common support region resulted in dropping 19 programme supported and 9 programme non-supported regions (outside of common support) from a further analysis, thus selecting a *comparable* 166 programme participants regions (out of a total of 185) and 120 programme non-participants regions (out of a total of 129) as relevant counterparts. In the next step the balancing property tests (t-test) were carried out to verify statistically the comparability of selected groups of regions in terms of observable covariates (Table 6).

The above tests show that the applied matching procedure (i.e. minimization of the standardized selection bias using calliper matching 0.21) considerably improved *comparability* of both groups of regions, making a counterfactual analysis more realistic. Indeed, previously existing significant differences (measured in terms of t-test) in variables between the group of regions supported from the SAPARD programme (D=1) and non-supported regions (D=0) *before* matching dropped *after* matching (differences became no more significant). This applies to all important variables determining both programme participation and outcomes, e.g. RDI 2002; unemployment rate, rural unemployment,

22 The standardized bias is the difference of the sample means in the treated and non-treated (full or matched) sub-samples as a percentage of the square root of the average of the sample variances in the treated and non-treated groups (Rosenbaum and Rubin, 1985).

23 This does not apply to local linear weighting function matching which first smoothes the outcome and then performs nearest neighbor matching. In this case more controls are used to calculate the counterfactual outcome than the nearest neighbor only (Leuven and Sianesi, 2007).

Table 6: Poland: Variables' balancing test between selected (common support region; calliper matching 0.21) programme supported and non-supported NUTS-4 regions (programme intensity per region)

		Mean		%reduct		t-test	
Variable	**Sample**	**Treated**	**Control**	**%bias**	**bias**	**t**	**p> t**
f1	Unmatched	-.14083	.32879	-50.9		-4.39	0.000
	Matched	-.03198	-.01991	-1.3	97.4	2.15	0.032
f2	Unmatched	-1.0652	-1.0738	4.5		0.39	0.698
	Matched	-1.0795	-1.1063	13.8	-209.0	0.58	0.560
f3	Unmatched	-.02088	.10599	-12.8		-1.14	0.254
	Matched	-.00283	.05661	-6.0	53.1	0.45	0.652
f4	Unmatched	-.11988	.06443	-19.1		-1.78	0.077
	Matched	-.10958	-.17486	6.8	64.6	0.63	0.532
f5	Unmatched	-.15212	.19606	-34.8		-2.95	0.003
	Matched	-.0.377	.07968	-11.7	66.3	0.52	0.605
f6	Unmatched	.37727	.3958	-14.5		-1.27	0.205
	Matched	.38806	.37875	7.3	49.8	0.83	0.407
f7	Unmatched	-.04476	.139	-17.1		-1.59	0.113
	Matched	-.014	-.0544	3.8	78.0	0.49	0.628
f8	Unmatched	.16594	-.20835	40.0		3.45	0.001
	Matched	.02705	-.04897	8.1	79.7	-1.43	0.153
f9	Unmatched	-.06558	.11431	-17.5		-1.58	0.114
	Matched	-.07627	-.02007	-5.5	68.8	0.36	0.722
f10	Unmatched	-.09891	-.00915	-9.1		-0.82	0.412
	Matched	-.06256	.09627	-16.2	-77.0	-0.19	0.852
f11	Unmatched	-.12477	.14517	-28.0		-2.51	0.013
	Matched	-.13269	-.17555	4.4	84.1	1.29	0.199
f12	Unmatched	.05226	.15766	-11.5		-1.01	0.313
	Matched	.07588	.07544	0.0	99.6	0.34	0.735
f13	Unmatched	.35259	-.14791	51.8		4.61	0.000
	Matched	.25845	.21174	4.8	90.7	-2.27	0.024
f14	Unmatched	-.97069	-1.2232	37.1		3.36	0.001
	Matched	-.97899	-1.0644	12.6	66.2	-0.79	0.430
f15	Unmatched	-.07538	.03328	-11.2		-0.99	0.322
	Matched	-.0608	-.07323	1.3	88.6	0.20	0.840
f16	Unmatched	.03438	-.10239	13.7		1.21	0.227
	Matched	.01498	-.03325	4.8	64.7	-0.69	0.493
f17	Unmatched	.34535	.11288	21.9		1.94	0.053
	Matched	.32236	.10717	20.3	7.4	-0.39	0.694
RDI2002	Unmatched	.01327	.03078	-19.6		-1.79	0.074
	Matched	.01506	.00871	7.1	63.7	0.63	0.528
unemploy2002	Unmatched	.09544	.09953	-13.9		-1.22	0.223
	Matched	.09601	.09504	3.3	76.1	0.84	0.404
unemplrur02	Unmatched	.6249	.52123	63.3		5.44	0.000
	Matched	.59893	.58412	9.0	85.7	-2.42	0.016

as well as factors f4 (the percentage of highest income groups and housing availability), f5 (subsidies and social expenditures), f8 (gas supply system) or f11 (Heating energy sector <pollution> and deaths), and others. Also other tests, e.g. pseudo R^2 (pseudo R^2 = 0.24 before matching and pseudo R^2= 0. 07 after matching) confirmed the high quality of the selected matching procedure and thus applicability of the used approach.

8.5. Calculation of policy evaluation parameters (ATT, ATE, ATU)

Comprehensive assessment of programme impact at a regional level requires separation of various important programme effects, e.g. effect on regions which participated in a given programme (Average Treatment Effect on the Treated - ATT); effect on an average region randomly selected from the pool of programme participants and non-participants (Average Treatment Effect – ATE) or an effect of the programme on the regions that did not participate (Average Treatment Effect on the Untreated – ATU).

In our study, the above policy evaluation parameters (ATT, ATE, ATU) were calculated on the basis of estimated propensity scores using the following programme impact indicators:

a. RDI

b. Unemployment rate (general)

c. Rural unemployment (percentage of rural unemployment in total unemployment)

The results of ATT, ATE and ATU calculations are shown in Table 7. Given these parameters the programme impact is quantified using a *conditional DID estimator*, i.e. combining PSM (ATT, ATE, ATU) and difference in differences (DID) methods.

8.6. Combined PSM and DID estimator

The application of the binary PSM method (including thresholds), and the conditional DID estimator to the assessment of the programme impact shows that the effect of the SAPARD programme (Measure 3) on the overall level of rural development in regions that participated in the programme (less developed regions) was almost negligible. Indeed, probably due to a low programme intensity and a short time horizon, the estimated impact of infrastructural measures (Measure 3) on the overall RDI in regions that participated in the programme (i.e. a difference between ATT in 2002 and ATT in 2005) was close to zero (the difference between the RDI in regions participating in the programme and regions non-supported remained almost constant over the years 2002-2005).

In contrast, a slight positive impact of SAPARD (Measure 3) was found on rural unemployment. When measured in absolute values, between 2002 and 2005 rural unemployment stayed on average (all 314 regions) at a similar level (approximately 58% of total unemployment). Yet, during the same period in our *comparable* groups (matched regions supported by the programme and similar control group) rural unemployment increased, due to *negative* economic conditions characterising these regions. Interestingly, in the same time period rural unemployment in the control group of regions (non-participants) grew stronger (0.0095) compared with the group of programme participants (0.0061). Consequently, the estimated ATT dropped from 0.0148 in 2002 (difference between 0.599 for D=1 and 0.584 for D=0) to 0.0114 in 2005 (difference between 0.605 for D=1 and 0.594 for D=0) thus indicating a slight but *positive*[24] impact of SAPARD (Measure 3) on rural unemployment in those regions supported by the programme.

24 Due to a negative context of the impact indicator (unemployment) a positive change in ATT (difference between after and before) would indicate a negative impact of the programme.

Table 7: Estimated policy evaluation parameters (per region basis)

Calculation basis	RDI			Rural unemployment		
	2002	2005	DID (2005 - 2002)	2002	2005	DID (2005 - 2002)
Unmatched 1 (185)	.01326	.0103	-.003	.6249	.6303	.0054
Unmatched 0 (129)	.03077	.0293	-.0015	.5212	.5309	.0097
Ø (314)	.0204	.0181	-.0023	.5823	.5894	.0071
Difference (1-0)	-.0175	-0.019	-.0015	.1036	.0993	-.0043
Difference (1- Ø)	-.00714	-.0078	-.00066	.0426	.0409	-.0017
Matched M 1 (166)	.0150	.0120	-.003	.5989	.6050	.0061
Matched M 0 (120)	.0087	.0056	-.003	.5841	.5936	.0095
ATT	.0063	.0063	0	.0148	.0114	-.0034
ATU	-.0097	-.0116	-.0019	.0162	.0139	-.0023
ATE	-.0003	-.0011	-.0008	.0153	.0125	-.0028

Furthermore, we found that the SAPARD programme (Measure 3) would have a slight but positive impact on rural unemployment (i.e. decrease) both in:

a. those regions that were previously excluded from the programme (a negative change in estimated ATU between 2002 and 2005), as well as

b. any other region randomly selected from a total sample of both groups of regions (a negative difference in ATE between 2002 and 2005).

From the policy point of view conclusions based on ATT parameters are especially important (i.e. impact on those regions which were supported from the programme). Concerning the conclusions (a) and (b) their relevance is restricted due to the fact that they include the effect on regions **j** for which the programme *was never intended/ designed* (from an administrative point of view these regions may be even programme ineligible).

8.7. Other programme intensity and participation criteria

As the measurement of the intensity of a region's participation in the SAPARD programme (Measure 3) on a *per region* basis (with a threshold) may appear problematic, especially in the case of small regions, two other alternative participation measures were applied:

1. programme exposure *per capita;* and

2. programme exposure *per km*

As in the case of programme exposure *per region* respective participation thresholds were set at the level of 66% of the country's average.

8.7.1. Intensity to programme exposure measured per capita and km basis

Use of other alternative measures of the intensity of programme participation (SAPARD funds under Measure 3 per capita or km^2)

combined with the application of the above thresholds (D=0 if regional programme exposure is below a 66% of regions' average; otherwise D=1) resulted in the following division of 314 NUTS-4 regions:

1. per capita: 188 regions supported and 126 regions non-supported, or

2. per km²: 178 regions supported and 136 region non-supported

As in the case of "programme exposure per region", in both settings (i.e. per capita and per km²) supported and non-supported regions were found to differ considerably in economic, social and environmental aspects of rural development. For example, when measuring the intensity of programme participation on per capita basis, the RDI (2002) in the group of regions supported from the programme (D=1) was as much as 2/3 lower compared with the group of programme non-participants (D=0); rural unemployment in 2002 in group 1 was much higher than in group 2 (64% compared with 48%); endowment with factor 4 (high income groups and housing availability) in group 1 was far below the country average (-0.18), whereas in group 2 it was far above (+0.16). Similar differences were also observable in the case of other partial indicators.

Clearly, significant differences between both groups of regions in terms of individual regional characteristics (RDI, factor endowments, etc.) confirm (similarly to the case of "programme intensity per region") the existence of a considerable **selection bias** preventing a direct comparability of both regional clusters within a counterfactual analysis.

8.7.1.1. Selection of appropriate matching algorithm

As with programme exposure on a *per region basis,* the selection of the best matching algorithm on a *per capita or km* basis was carried out using the method described in Section 8.3.2.3. The application of the above technique resulted in the selection of a radius calliper 0.23 (for *per capita* setting); and Gaussian kernel (bandwidth 0.14) (for *per km* setting) as the matching algorithms that guaranteed the minimization of a standardized bias (after matching).

Table 8: Poland: Variables' balancing test between selected (common support region; caliper matching 0.23) programme supported and non-supported NUTS-4 regions (programme intensity per capita basis)

		Mean		%reduct		t-test	
Variable	**Sample**	**Treated**	**Control**	**%bias**	**bias**	**t**	**p> t**
f1	Unmatched	-.20661	.43812	-72.7		-6.18	0.000
	Matched	.00017	.14994	-16.9	76.8	1.32	0.188
f2	Unmatched	-1.0447	-1.1046	31.0		2.70	0.007
	Matched	-1.0619	-1.0619	3.8	87.8	-0.46	0.645
f3	Unmatched	-.15312	.30633	-46.5		-4.23	0.000
	Matched	-.15905	-.10753	-5.2	88.8	1.98	0.049
f4	Unmatched	-.18231	.16197	-35.4		-3.35	0.001
	Matched	-.14409	-.16485	2.1	94.0	1.46	0.145
f5	Unmatched	-.11573	.15006	-26.2		-2.23	0.027
	Matched	-.01357	-.04807	3.4	87.0	0.41	0.681
f6	Unmatched	.36812	.40989	-33.5		-2.88	0.004
	Matched	.3778	.3763	1.2	96.4	1.50	0.135
f7	Unmatched	-.05652	.16092	-20.5		-1.88	0.062

	Matched	-.00996	-.5025	3.8	81.5	0.94	0.348
f8	Unmatched	.122257	-.15255	29.7		2.51	0.013
	Matched	-.08669	-.02593	-6.6	77.9	-0.88	0.378
f9	Unmatched	-.05054	.09616	-14.1		-1.29	0.199
	Matched	-.0581	.0301	-8.5	39.9	0.94	0.346
f10	Unmatched	-.09781	-.00865	-9.1		-0.81	0.417
	Matched	-.04331	-.03939	-0.4	95.6	0.36	0.719
f11	Unmatched	-.10092	.11601	-22.6		-2.00	0.046
	Matched	-.12575	-.07475	-5.3	76.5	1.13	0.260
f12	Unmatched	.08332	.11381	-3.3		-0.29	0.771
	Matched	.11579	.1621	-5.0	-51.9	0.64	0.524
f13	Unmatched	.08814	.23476	-14.9		-1.31	0.193
	Matched	.07557	.06029	1.6	89.6	0.42	0.673
f14	Unmatched	-1.1123	-1.0179	-14.2		-1.23	0.219
	Matched	-1.0672	-1.0502	-2.6	82.0	0.56	0.576
f15	Unmatched	-.1193	.10141	-23.3		-2.02	0.045
	Matched	-.05125	-.00432	-4.9	78.7	1.23	0.219
f16	Unmatched	.10282	-.20776	31.9		2.76	0.006
	Matched	.01045	-.15683	17.2	46.1	-0.53	0.593
f17	Unmatched	.27198	.21682	5.1		0.46	0.649
	Matched	.28487	.31151	-2.5	51.7	0.02	0.984
RDI2002	Unmatched	.00775	.03943	-35.1		-3.27	0.001
	Matched	.01319	.01344	-0.3	99.2	1.25	0.213
unemploy2002	Unmatched	.09759	.09641	4.0		0.35	0.727
	Matched	.0982	.10044	-7.6	-90.7	-0.56	0.574
unemplrur02	Unmatched	.6449	.48892	100.2		8.69	0.000
	Matched	.60829	.58422	15.5	84.6	-2.12	0.035

Table 9: Poland: Variables' balancing test between selected (common support region; kernel (Gaussian) matching bw 0.14) programme supported and non-supported NUTS-4 regions (programme intensity per km² basis)

		Mean		%reduct		t-test	
Variable	**Sample**	**Treated**	**Control**	**%bias**	**bias**	**t**	**p> t**
f1	Unmatched	-.15961	.32919	-52.7		-4.62	0.000
	Matched	-.09412	.02708	-13.1	75.2	1.46	0.144
f2	Unmatched	-1.0721	-1.0643	-4.0		-0.35	0.726
	Matched	-1.075	-1.0922	8.9	-121.4	0.12	0.902
f3	Unmatched	.17722	-.15981	35.8		3.10	0.002
	Matched	.12844	.13437	-0.6	98.2	-1.35	0.177
f4	Unmatched	-.05723	-.02705	-3.2		-0.29	0.771
	Matched	-.03782	-.00818	-3.2	1.8	-0.11	0.911
f5	Unmatched	-.02007	.00532	-2.4		-0.21	0.831
	Matched	-.02836	.07239	-9.7	-296.8	-0.08	0.037
f6	Unmatched	.38165	.3891	-5.9		-0.51	0.608

	Matched	.38469	.38244	1.8	69.7	-0.03	0.976
f7	Unmatched	-.05031	.13681	-17.7		-1.63	0.104
	Matched	-.02258	-.03908	1.6	91.2	0.30	0.762
f8	Unmatched	.13201	-.14468	29.5		2.55	0.011
	Matched	-.0801	-.20188	13.0	56.0	0.21	0.833
f9	Unmatched	-.10024	.15042	-24.5		-2.23	0.026
	Matched	-.10516	-.11022	0.5	98.0	0.63	0.531
f10	Unmatched	-.14401	.04526	-19.8		-1.75	0.081
	Matched	-.10289	-.02376	-8.3	58.2	0.46	0.648
f11	Unmatched	-.23425	.27457	-54.8		-4.90	0.000
	Matched	-.19026	-.12763	-6.7	87.7	2.18	0.030
f12	Unmatched	.0934	.09838	-0.5		-0.05	0.962
	Matched	.07728	.03162	5.0	-816.6	0.51	0.612
f13	Unmatched	.29329	-.04454	34.6		3.08	0.002
	Matched	.23664	.21985	1.7	95.0	-1.03	0.303
f14	Unmatched	-1.1065	-1.0325	-11.2		-0.98	0.330
	Matched	-1.096	-1.095	-0.2	98.6	0.63	0.528
f15	Unmatched	-.08164	0.3589	-12.2		-1.08	0.281
	Matched	-.05906	-.00934	-5.2	57.7	0.18	0.856
f16	Unmatched	.06935	-.14112	21.3		1.88	0.061
	Matched	.02832	-.09583	12.6	41.0	-0.08	0.935
f17	Unmatched	.22286	.28516	-5.9		-0.52	0.603
	Matched	.22758	.15756	6.6	-12.4	0.35	0.724
RDI2002	Unmatched	.02019	.02082	-0.7		-0.07	0.948
	Matched	.02083	.02144	-0.7	5.4	-0.06	0.951
unemploy2002	Unmatched	.09263	.10299	-35.7		-3.15	0.002
	Matched	.09336	.09449	-3.9	89.1	1.10	0.272
unemplrur02	Unmatched	.60861	.54789	35.7		3.12	0.002
	Matched	.59535	.55396	24.4	31.8	0.07	0.944

The application of the above matching algorithms led to a significant improvement of balancing properties between selected covariates in both settings (Tables 8 and 9) and thus a better comparability between the group of regions supported from the programme with a control group of regions (non-supported regions).

8.7.1.2. Combined PSM and ATT estimator (conditional DID estimator)

The application of the conditional DID estimator to a measurement of the programme impact at regional level (using programme intensity *per capita* basis as a criterion for programme participation) during the period 2002-20005 shows, as in the case of the *per region* indicator, an almost negligible effect of the SAPARD programme (Measure 3) on the overall quality of life (DID in ATT = -0.0011) and rural unemployment (DID in ATT = 0.0006) in regions supported from the programme (see Table 10). These results differ from results obtained by applying traditional evaluation techniques (e.g. DID using a group of non-participants or the country average as respective controls), which showed a positive effect on RDI (i.e. 0.0069 or 0.0077) and a slightly positive impact on rural unemployment (i.e. -0.0007 and -0.0003).

Table 10: Poland: Estimated policy evaluation parameters (per capita basis; M3 per capita)

Calculation basis	RDI			Rural unemployment		
	2002	2005	D I D (2005 - 2002)	2002	2005	D I D (2005 - 2002)
Unmatched 1 ( )	.0077	.0026	-.0051	.6449	.6517	.0068
Unmatched 0 ()	.0394	.0412	.0018	.4889	.4965	.0076
Ø (314)	.0971	.0843	-.0128	.5823	.5894	.0071
Difference (1-0)	-.0316	-.0385	.0069	.1559	.1552	-.0007
Difference (1-Ø)	-.0894	-.0817	.0077	.0626	.0623	-.0003
Matched M 1 ()	.0131	.0080	-.0051	.6082	.6157	.0075
Matched M 0 ()	.0134	.0094	-.0040	.5842	.5910	.0068
ATT	-.0002	-.0013	-.0011	.0240	.0246	.0006
ATU	-.0112	-.0134	-.0022	.0521	.0559	.0038
ATE	-.0051	-.0067	-.0016	.0364	.0385	.0021

8.8. Sensitivity of obtained results

The sensitivity of obtained results was estimated using the procedure proposed in Rosenbaum (2002). The approach allows the determination of how much hidden bias would need to be present to render the null hypothesis of no effect, or in another words, how strongly an unmeasured variable must influence the selection process in order to undermine the implications of a standard (binary) propensity score matching analysis (Caliendo and Kopeinig, 2005).

The procedure applied in this study calculates Rosenbaum bounds for average treatment effects on the programme supported regions in the presence of unobserved heterogeneity (hidden bias) between treatment and control cases[25].

In the case of a *per region* basis, sensitivity analysis shows that the estimated positive effect of SAPARD (Measure 3) on rural unemployment is rather sensitive to unobservable heterogeneity (i.e. sensitive to possible deviations from the identifying unconfoundedness assumption). Indeed, an increase of gamma by 10% to Γ = 1.1 would result in insignificance of obtained results at the 10% significance level (sig + = 0.14 in 2002 and sig + = 0.16 in 2005). Of course, this result does not mean that unobserved heterogeneity exists and there is no effect of the SAPARD programme (Measure 3) on rural unemployment. This result only states that the confidence interval for the effect would include zero if an unobservable variable caused the odds ratio between regions supported from the programme and the control group to be higher than 1.1. In the case of *per capita* and *per km* , estimated results are less sensitive, i.e. only a hidden bias increasing gamma to 1.2 (1.4) would lead to an insignificance of obtained results.

25 The procedure calculates Wilcoxon signrank tests that give upper and lower bound estimates of significance levels at given levels of hidden bias. Under the assumption of additive treatment effects, rbounds also provides Hodges-Lehmann point estimates and confidence intervals for the average treatment effect on the treated (Gangl, M., in STATA 10.1; 2007).

8.9. Application of a generalized propensity score matching to the assessment of SAPARD's impact at regional level

An important problem linked to the evaluation of programme impact using the binary PSM method, in a situation where almost all regions received a support from the given programme, is the small size or a non-availability of a control group (D=0). Depending on data, this problem can be partly solved within a framework of binary treatment (i.e. using the binary PSM method) by applying a threshold and considering regions experiencing low programme intensity (below the threshold) as programme *non-supported regions* (see Chapter 8.3 above).

However, beyond some uncertainties as to the appropriateness of a given threshold level, the application of the "threshold approach" in combination with a traditional (i.e. binary) PSM method to the assessment of programme impact, may also not be particularly *efficient* as this framework largely disregards information normally available about the programme intensity (measured per *region, per capita or km* basis). Indeed, in order to learn more about the effectiveness of a given programme's dependence on the level of programme exposure (effectiveness dynamics) a more sophisticated approach has to be applied. If the level of programme support (i.e. exposure to programme participation) is a continuous variable (e.g. programme financial allocation by *regions, per capita or per km*) a *generalized propensity score matching* (GPSM) methodology is especially advantageous. Especially interesting here is the possibility of the estimation of the average and marginal potential outcomes that correspond to specific values of continuous programme doses (i.e. *for each level of programme support*) by means of a dose-response and derivative dose-response functions.

Application of the GPSM methodology to an analysis of the impact of the SAPARD programme (Measure 3) in Poland was carried out using information on a *per region* basis as a respective measure of programme intensity. The analytical steps are described in Chapter 4.7.

8.9.1. Estimation of GPS and dose response function

Given that for each region **i** we observe a vector of specific regional covariates (X= f1-f17), the level of support (T) from the SAPARD programme (Measure 3), and the potential outcome corresponding to a given programme intensity level (Y(T)= RDI, rural unemployment, etc.) our basic objective is to estimate the average and the derivative of the dose-response function (ADRF) = μ(t) and DDRF = v (t), where:

μ(t) = E [Y(T)] = the average effect of the programme in dependence on programme intensity;

v(t) = E[Y(T+1)- Y(T)] = derivative dose response function, in dependence on programme intensity.

As shown in Hirano and Imbens (2004) the conditional density of the treatment given covariates, the Generalized Propensity Score (GPS) has a balancing property similar to the balancing property of the propensity score for binary treatments. Adjusting for the GPS therefore removes all bias associated with differences in region specific covariates.

In our study, region specific GPS was estimated as a conditional density of treatment (T) given covariates describing individual characteristics of the region (factors f1-f17). The parameters of the treatment function (conditional distribution of treatment) were estimated using maximum likelihood.

The major steps and results of generalized propensity score estimation (programme intensity *per region*) are described below.

8.9.1.1. Estimation of the treatment function

The conditional distribution of support intensity (treatment function) given region specific

Table 11: Poland Results of treatment function estimation (version: per region)

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
eq1						
f1	-.2168733	.0313629	-6.91	0.000	-.2783434	-.1554033
f2	-.1178106	.1921511	-0.61	0.540	-.49442	.2587987
f3	-.0693768	.031529	-2.20	0.028	-.1311724	-.0075811
f4	-.0544577	.0329246	-1.65	0.098	-.1189888	.0100733
f5	-.1062281	.0312695	-3.40	0.001	-.1675152	-.0449409
f6	-.2075365	.2672592	-0.78	0.437	-.7313549	.3162819
f7	-.0179871	.0294356	-0.61	0.541	-.0756798	.0397055
f8	.1777737	.0305866	5.81	0.000	.117825	.2377224
f9	-.0840799	.0291012	-2.89	0.004	-.1411172	-.0270425
f10	-.0247202	.0309118	-0.80	0.424	-.0853063	.0358658
f11	-.1219499	.0322142	-3.79	0.000	-.1850886	-.0588111
f12	-.0357839	.0322489	-1.11	0.267	-.0989906	.0274227
f13	.2086801	.0312673	6.67	0.000	.1473972	.2699629
f14	.0410695	.0507565	0.81	0.418	-.0584114	.1405503
f15	-.0337561	.0306387	-1.10	0.271	-.0938068	.0262946
f16	.0463776	.0304062	1.53	0.127	-.0132174	.1059726
f17	.1091553	.0297322	3.67	0.000	.0508813	.1674293
_cons	1.734603	.2834282	6.12	0.000	1.179094	2.290112
eq2						
_cons	.5078574	.0202657	25.06	0.000	.4681374	.5475775

Logistic regression	**Number of obs**	**Wald chi2(17)**	**Prob > chi2**	**Log likelihood**
	= 314	= 235.07	= 0.0000	= -232.794

covariates was estimated on the basis of the zero-skewness log transformation function with factors f1-f17 as function arguments, and the programme intensity level per region as a dependent variable. The treatment function was estimated by applying the maximum likelihood estimator to eq 14. Results of the estimation are shown in Table 11.

In the next step, normality assumptions of the estimated function were tested. Test for normality of the disturbances (STATA skewness and kurtosis test for normality) confirmed that the assumption of normality was statistically satisfied at .05 level (Table 12).

Table 12: Poland: Results of skewness/kurtosis test for normality of the disturbances (version: per region)

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
res_etreat	314	0.253	0.480	1.81	0.4038

8.9.1.2. Calculation of the GPS and testing the balancing property

Given region specific information on Ti, Xi as well as estimated under 8.9.1.1. parameters ($\hat{\gamma}$ and $\hat{\sigma}^2$ the value of the GPS was calculated (evaluated) for each region according to eq 15.

Having estimated the GPS, similar to the case of binary treatment, it is crucial to investigate whether the GPS specification is adequate, i.e. whether it balances the covariates (Hirano and Imbens, 2004; Bia and Mattei, 2007; Kluve, et al. 2007). In order to implement the balancing property tests we divided the range of programme intensity into four treatment intervals (i.e. less than 5m PLN per regions; 5-10m per region; 10-20m per region; 20-43m per region), with 169 observations in the first group, 90 in the second, 45 in the third, and 10 in the last treatment interval. Respective tests were carried out on the conditional mean of the pre-treatment variables given the generalized propensity score is not different between regions that belong to a particular treatment interval and regions that belong to all other treatment intervals. The balancing tests were performed for each single variable included in the list of covariates and each mean treatment interval.

According to a standard two-sided t-test we found that in all treatment intervals the balancing property was satisfied at a level lower than 0.01, thus the covariates in both groups of regions were not significantly different (t-test for each of the 17 covariates and each four groups of intervals are shown in Appendix 1).

8.9.2. Modelling the conditional expectation of the programme outcome

Given Ti and the estimated GPS (R_i) for each NUTS-4 regions, the conditional expectation of the programme outcome measured in terms of RDI ($Y = \Delta$ RDI) was modelled as a flexible function of its two arguments (T_i and R_i) according to eq 17 (polynomial quadratic function).

$$Y = b_0 + b_1T + b_2T^2 + b_3GPS + b_4GPS^2 + b_5T*GPS \qquad (17)$$

The results of this estimation, with the outcome variable representing the change of the overall level of rural development (i.e. RDI2005-RDI2002) and Ti, Ti square, Ri and Ri square as independent variables are shown in Table 13.

As shown in Hirano and Imbens (2004) in this model the estimated coefficients do not have

Table 13: Poland: Estimated parameters of the conditional expectation of the programme outcome (SAPARD programme – Measure 3)

The regression model 1s: Y = T + T^2 + GPS + GPS^2 + T*GPS				
Source	ss	df	MS	Number of obs = 314
Model	.002798331	5	.000559666	F(5, 308) = 1.12
Residual	.153406037	308	.000498072	Pr ob > F = 0.3477
Total	.156204369	313	.000499055	R-squared = 0.0179 Adj R-squared = 0.0020 Root MSE = .02232

Δ RDI (2002-2005)	Coef.	Std. Err.	t	P > t	[95% Conf.	Interval]
b_1	-.0015462	.0008133	-1.90	0.058	-.0031466	.0000542
b_2	.0000438	.0000217	2.02	0.044	1.19e-06	.0000865
b_3	-.0129297	.0300552	-0.43	0.667	-.0720691	.0462098
b_4	.0122018	.0304798	0.40	0.689	-.0477732	.0721769
b_5	.0006998	.0010106	0.69	0.489	-.0012887	.0026883
_const	.0047889	.0068562	0.70	0.485	-.0087019	.0182798

a causal interpretation. Yet, the parameters of the estimated regression model (17) are later used to estimate the outcome of programme support in particular at level T.

8.9.3. Estimation the average potential outcome for each level of treatment (entire dose-response function)

Given the estimated individual conditional expectations of the programme outcome at the individual (regional) programme intensity levels, the entire dose-response function (DRF) was computed as the *average* potential outcome for each level of treatment according to eq. 16.

After averaging the dose-response over propensity score for each level of T, the marginal causal effects were computed in the form of the *derivative* dose-response function E[Y(T+1) - Y(T)].

In our study bootstrapping methods were used to obtain standard errors that take into consideration the estimation of GPS and parameters of the estimated conditional expectation function.

8.10. Impact of SAPARD programme (Measure 3) on the overall level of rural development

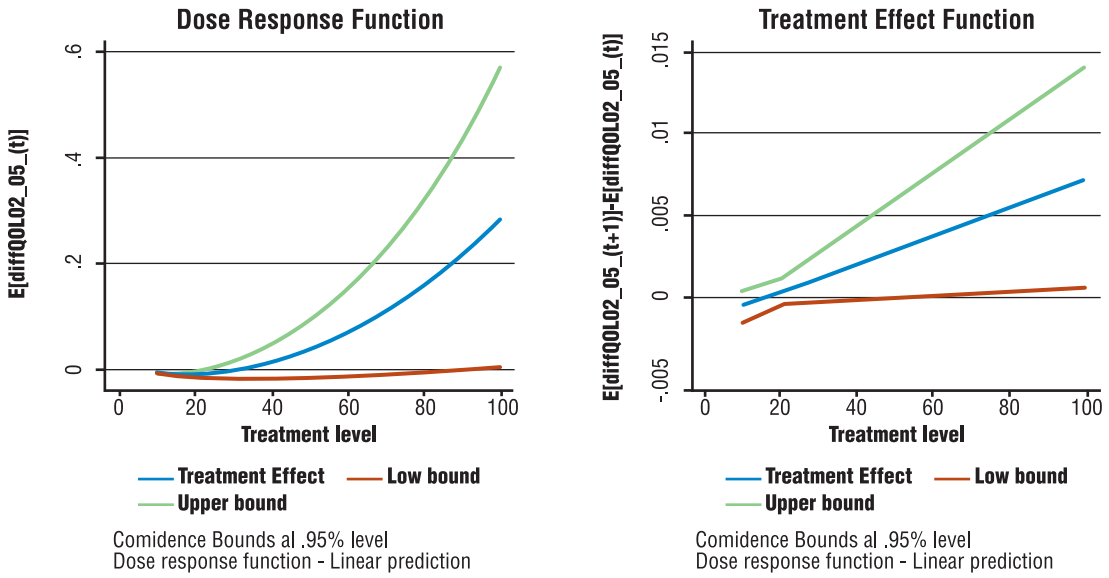
The results of the above calculations, together with the estimates of the *derivative* dose-response function that provides information about the *marginal* effects of the SAPARD programme (Measure 3) on the overall level of rural development (measured in terms of the RDI) are shown in Table 14. A graphical presentation of obtained results (i.e. impact of the SAPARD programme (Measure 3) on the overall level of rural development) is shown in Figure 4.

The application of the GPS matching and the dose response function to the assessment of the impact of the SAPARD programme (Measure 3) on the overall level of rural development in regions supported by the programme (a change of the RDI as an impact indicator), enables a more precise estimation of the effects of the SAPARD programme when compared with traditional evaluation techniques or methodologies based on binary PSM methods.

Table 14: Poland: Estimated effects of SAPARD (Measure 3) on the overall level of rural development (RDI) by means of dose-response and derivative of dose-response functions.

T_level	T_level_plus	dose_response	diff_dose_response	se_dose_response_bs	se_diff_dose_response_bs
10	11		-.0061104	-.0005007	.002162
20	21	-.0080311	.0002364	.0048819	.0004013
30	31	-.0017626	.0011096	.0075473	.0007127
40	41	.013311	.0019934	.0143247	.0010896
50	51	.0372164	.0028754	.0258549	.001476
60	61	.0699312	.0037552	.0417073	.0018652
70	71	.1114378	.0046338	.0616703	.0022558
80	81	.161726	.0055116	.0856593	.0026473
90	91	.2207902	.006389	.1136371	.0030395
100	101	.2886275	.0072662	.1455857	.0034323

Figure 4: Poland: Estimated dose response function, treatment effect function and 95% confidence bands for the impact of SAPARD programme (Measure 3) on the RDI (criterion: per region) in years 2002-2005



The main findings from the application of the GPS matching and DRF are as follows:

1. Results from the GPS and dose response function generally show a *positive* effect of SAPARD (Measure 3) on the overall level of rural development in supported regions. However, they also show that this positive impact was observable only for regions supported from the programme at a higher intensity level (i.e. above approximately 17m PLN per region). Negligible programme effects were mainly found in regions with a *low* programme intensity (this only applies to regions that received less than 40% of the maximum support level, i.e. or lower than 80% of the average programme intensity).

2. An *increase* of the intensity of programme support (*per region* basis) was found to bring about a significant increase of returns (positive change in the overall level of rural development or the RDI).

3. The highest effects of the SAPARD programme (Measure 3) were found in those regions which received the highest programme support (i.e. regions which obtained from the programme between 20-43m PLN from the programme).

4. Not surprisingly, taking into consideration a generally low absolute level of programme support, the marginal effectiveness of SAPARD funds (Measure 3) was found to be highest in regions that received absolute support far above an average support level. This shows that an expected threshold of programme intensity (rural investments) causing *diminishing* returns was well *above* the obtained maximum (i.e. above 43m PLN per NUTS-4 region).

5. For some reason (probably due to high unit costs of the programme), the effectiveness of the SAPARD programme (Measure 3) in regions that received the smallest absolute support (i.e. less than €100k per region) appeared to be *negative*.

6. While the estimated dose response function shows a plausible causality between SAPARD funds (Measure 3) and the overall rural development, the estimated

95% confidence intervals were found to become wider together with the intensity of programme support, i.e. uncertainty increased (one reason could be a small number of data observations (=10) in the upper scale of support).

8.11. Impact of the SAPARD programme (Measure 3) on rural unemployment

Another important outcome (impact) indicator that may be used to assess the effects of the SAPARD programme in regions that received programme support is the change in rural unemployment. In principle, all steps to assess the impact of SAPARD on rural unemployment are similar to those carried out for the assessment of the programme on the overall level of rural development. The only difference is the selection of the outcome indicator (i.e. a change in rural unemployment ratio instead of a change in the RDI).

The results of the application of GPS and a dose response function methodology (including derivative dose-response function) to the evaluation of SAPARD impact on rural unemployment are shown in Table 15. Graphical results of SAPARD impact on rural unemployment are presented in Figure 5.

The main findings from the application of GPS matching and dose response (and derivative dose response) functions to the measurement of the effects of the SAPARD programme (Measure 3) on rural unemployment in Poland (years 2002-2005) are as follows:

1. The SAPARD programme (Measure 3) was found to have a slight but positive effect on rural unemployment in NUTS-4 regions in Poland (years 2002-2005), i.e. rural unemployment was found to decrease slightly due to the SAPARD programme (the results of GPS were lower compared with the effects estimated by using a binary PSM method).

2. Also, as in the case of the RDI, the impact of the SAPARD programme (Measure 3) on rural unemployment was found to be highest in those regions that received the maximum programme support (above 20m PLN per region).

3. The impact of SAPARD (Measure 3) on rural unemployment in regions that received the lowest amount of funds from the programme was found to be almost zero (or negative).

4. With regard to marginal effects of the SAPARD programme on rural unemployment, these were positive at all programme intensity levels. Yet, the size of these effects was found to be relatively small and the estimated marginal effects remained almost constant along with the increase of programme intensity.

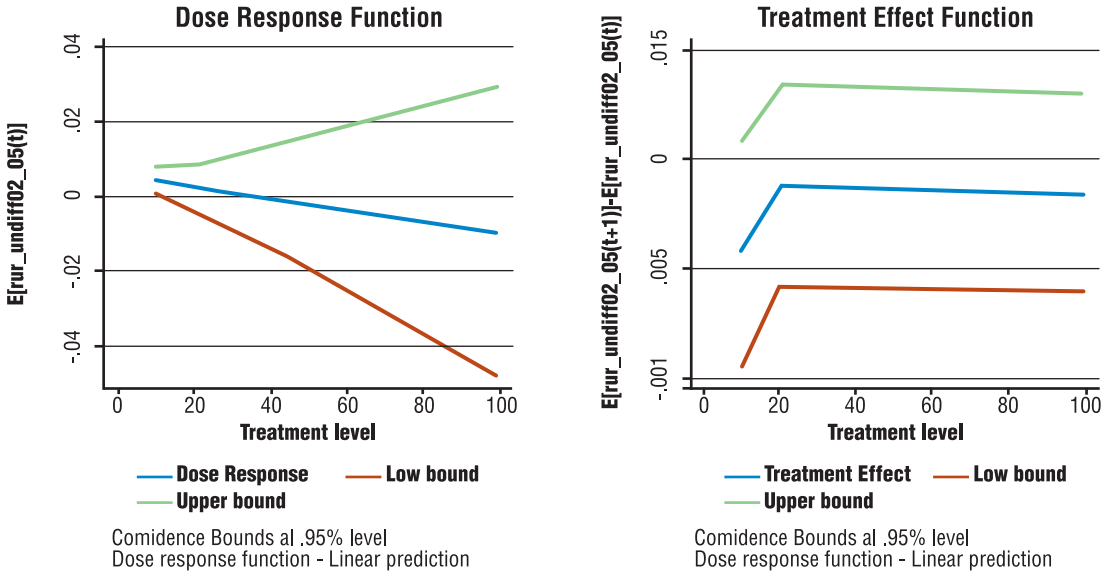
5. While the estimated dose response function shows a plausible causality between SAPARD funds (Measure 3) and the diminution of rural unemployment, the estimated 95% confidence intervals become wider along with the intensity of treatment (programme impacts become more uncertain).

Table 15: Poland: Estimated effects of SAPARD (Measure 3) on the rural unemployment by means of the dose-response function and the derivative of dose-response function

T_level	T_level_plus	dose_response	diff_dose_ response	se_dose_ response_bs	se_diff_dose _response_bs
10	11	.0046352	-.0004323	.0018734	.0002559
20	21	.0022044	-.000135	.0032044	.0002325
30	31	.0009392	-.0001275	.0047555	.0002362
40	41	-.0003952	-.0001403	.006669	.0002329
50	51	-.0018408	-.0001488	.0087418	.0002304
60	61	-.0033525	-.0001535	.0108872	.0002289
70	71	-.0048996	-.000156	.0130704	.000228
80	81	-.0064659	-.0001573	.0152758	.0002275
90	91	-.0080426	-.0001581	.0174956	.0002272

** = in case of unemployment a negative change in dose response function (or derivative dose response function) between years 2005 and 2002 indicates positive impacts of the programme.*

Figure 5: Poland: Estimated dose response function, treatment effect function and 95% confidence bands for the impact of SAPARD programme (Measure 3) on the rural unemployment (criterion: per region) in years 2002-2005



9. Assessment of the impact of the SAPARD programme in Slovakia

9.1. Rural Development Index as an impact outcome indicator

An important impact indicator applied to the assessment of the overall impact of the SAPARD programme in Slovakia is the Rural Development Index (RDI).

The RDI in Slovakia was calculated for all (72) rural NUTS-4 regions and the years 2002-2005 according to eq 7, on the basis of 337 regional indicators (21 region- and time-specific factors) and weights obtained from the estimated migration function. Territorial distribution of the RDI in Slovakia (by NUTS-4 regions) over the period 2002-2005 is shown in Figure 6 (below).

During the years 2002-2005, the estimated value of the RDI in Slovakia ranged from -0.51 to +0.91 (i.e. the regional discrepancies in the overall level of rural development were stronger in Slovakia than in Poland). As expected, the highest values of the RDI (i.e. highest development level of rural areas) were found in high performing regions located in West Slovakia (e.g. Senec, Pezinok, Dunajska Streda, Galanta, etc.). On the other hand the lowest RDI values (i.e. the lowest level of the overall rural development) were found in regions located in Eastern Slovakia and Central Slovakia (e.g. Gelnica, Stropkov, Namestovo, Kezmarok, Stara Lubovna).

The results obtained therefore confirm a clear typographic division of Slovakia into western, central and eastern sub-areas based on the performance of individual regions, and reiterate a general opinion that the level of rural development in Slovakia decreases considerably from West to East.

Figure 6: Distribution of RDI (by NUTS-4 regions) in years 2002-2005

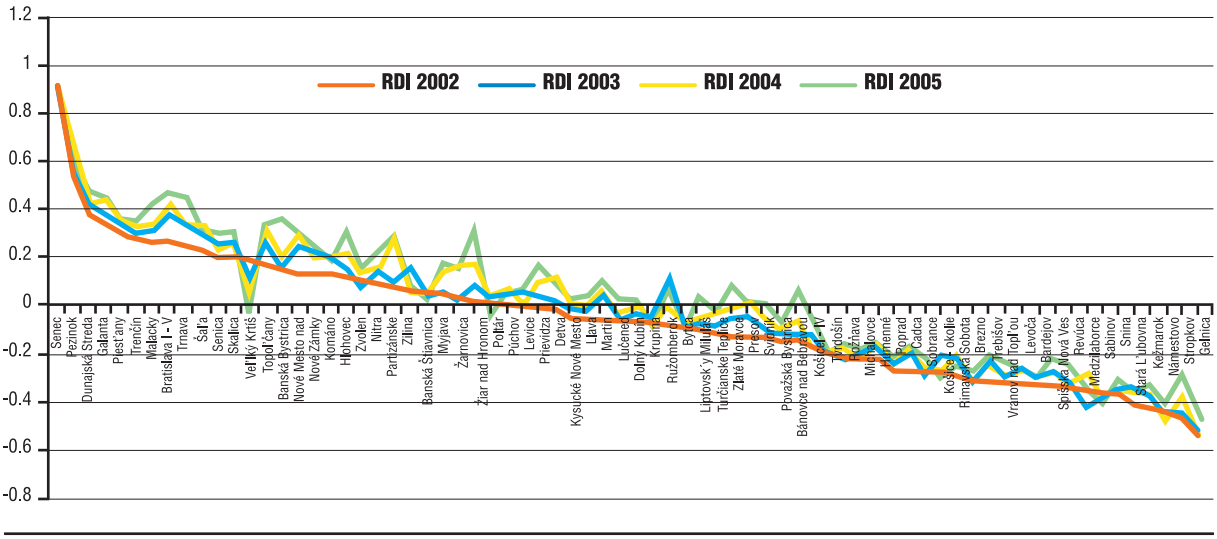
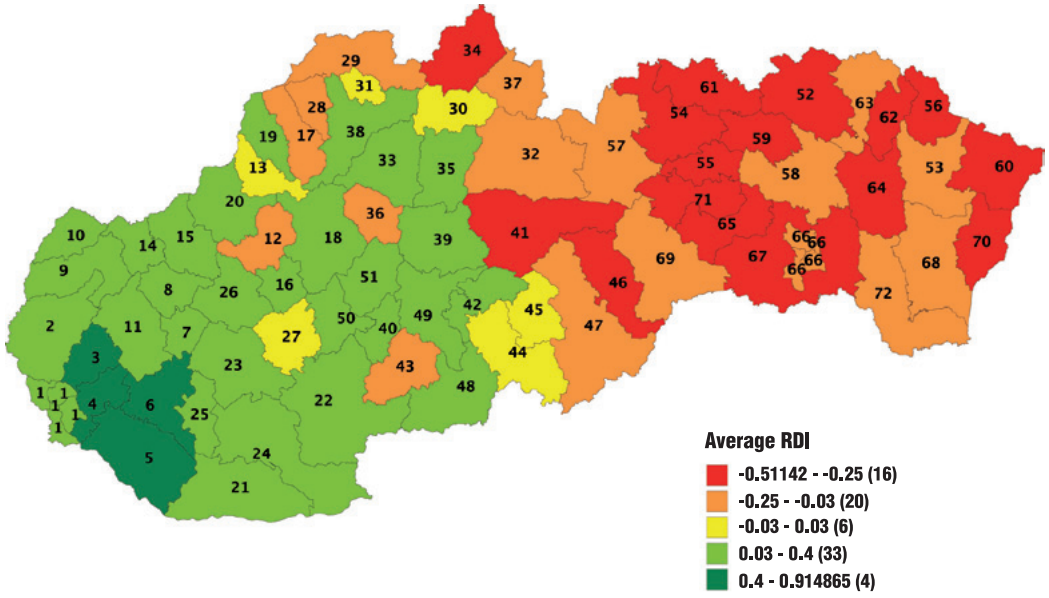


Figure 7: Distribution of RDI (average in years 2002-2005)



9.2. Regional characteristics as the main covariates determining selection of the region to the SAPARD programme

As mentioned before, the estimation of the RDI in Slovakia was carried out on the basis of 21 region- and time-specific *factors*, obtained by applying a principal component method to 337 regional specific coefficients describing various aspects of rural development. Application of the above methodology enabled the description of individual rural regions (a unique identification) in terms of their socio-economic and environmental characteristics (factors f1-f21). The overview of the main individual regional characteristics (factors) and their estimated social weights is shown in Table 16.

In the case of EU rural development programmes, the decision to select a particular region to a given structural or rural development programme is normally taken by a respective national Programme Managing Authority; this decision is made on the basis of strengths and weaknesses analysis (SWOT[26]). While SWOT analysis is a subjective assessment of a current situation in a given region, it draws upon regional data, including various partial socio-economic and environmental indicators. Here we apply a similar procedure in order to simulate a selection process of a given region to the SAPARD programme. I.e. by explaining a regional specificity and individual regional performance in terms of factors (f1-f21) we consider implicitly all important partial coefficients that are usually accounted for in a SWOT analysis.

Following this approach, factors (f1-f21), representing individual regional characteristics in the years 2002-2005, are used later (i.e. in estimation of a logit model or within a framework of generalized propensity score matching) as *the main covariates explaining differences in regional performance and the probability of selection of an individual region into a specific rural development programme (e.g. SAPARD programme)*.

26 SWOT is an acronym for Strengths, Weakness, Opportunities and Threads analysis.

Table 16: Slovakia: Individual rural development components and their social weights (2002-2005)

Factors	Rural development component	Estimated social weight
f1	Spatial density of social and retail infrastructure (per km²)	0.048
f2	Availability of social services and technical infrastructure (per capita)	-0.107
f3	Social conditions and living environment (incl. availability of dwelling)	0.096
f4	Agriculture and natural endowment	0.121
f5	Availability of young people's infrastructure (per capita)	0.015
f6	Spatial density of public utilities and social infrastructure: gas pipelines, water-supply-system (per km²)	0.044
f7	Density and structure of enterprises	-0.009
f8	Density of vocational secondary schools	-0.053
f9	Hotels and recreation facilities	0.014
f10	Endowment with special schools	-0.081
f11	Availability of social facilities (per capita)	-0.0002
f12	Accommodation endowment	0.036
f13	Public facilities	0.114
f14	Availability of retail infrastructure (per capita)	0.076
f15	Social facilities	0.031
f16	Primary schools	0.031
f17	Houses of social services	0.028
f18	Basic schools of art, etc.	0.003
f19	Density of specialized state secondary schools	-0.016
f20	High-standard tourist accommodations <negative loadings!>	-0.009
f21	Policlinics, grammar schools, sport grounds	0.038

9.3. Scope and distribution of funds from the SAPARD programme in Slovakia

Estimation of the impact of the SAPARD programme at regional basis requires information about regional distribution and intensity of total SAPARD funds (i.e. Measures 1, 2, 3, 4a, 4b, 5, 6, 7, 8, 9) between the years 2002-2004.

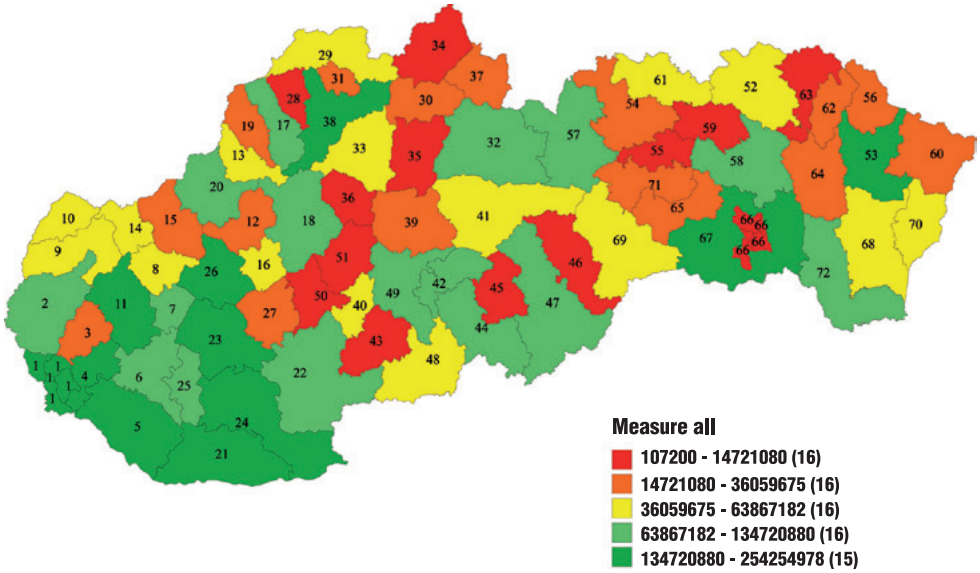
The implementation of the SAPARD programme in Slovakia resulted in the support of approximately 904 projects for a total amount of 4745m SKK funds (€111.5m). The majority of SAPARD funds were allocated to the Priority 1 "Improving of agricultural production sector including food industry" (61% of total SAPARD funds), e.g. investment projects in the agricultural and food industry sectors. This was followed by Priority 2 "Sustainable rural development" (38%), e.g. diversification and investment in rural infrastructure; agro-tourism projects or environmental projects. Funds for priority 3 "Development of human activities", e.g. technical assistance (1%), were allocated last.

A statistical analysis of SAPARD distribution by regions indicates a high variability of programme support intensity (see Table 17). While an average region (NUTS-4) in Slovakia received approx. 64.1m SKK from the SAPARD programme, some regions (e.g. Nitra region in West Slovakia) received more than 254m SKK (four times more than the country average). On the other hand, some other regions received only 0.1m SKK (€25k), e.g. the Poltar region in Middle Slovakia.

Table 17: Slovakia: Statistical distribution of SAPARD funds (by region)

Variable	Obs	Mean	Std. Dev.	Min	Max
Mall	72	6.41e+07	5.77e+07	107200	2.54e+08

Figure 8: Slovakia: Programme intensity (Measure 3) across regions



The geographical distribution of SAPARD funds (per region basis) across NUTS-4 regions shows that, contrary to some expectations, the majority of available programme resources were allocated to the *best developed regions* of West Slovakia (55% of total funds or 95.8m SKK per region) followed by Middle Slovakia (24% of total funds or 45.1m SKK per region) and Eastern Slovakia (21% of total funds or 44.9m SKK per region). A similar picture was obtained when considering SAPARD intensity on a per capita basis, i.e. the highest programme intensity was measured in the best developed regions located in West Slovakia (1135 SKK per capita =100%), followed by Middle Slovakia (69%) and Eastern Slovakia (65%). An analysis of the geographical distribution of SAPARD therefore shows, that programme funds were merely used to reinforce the market position of relatively well performing Slovak enterprises (i.e. mostly large agricultural farms and food industry companies) located in relatively well developed regions[27].

While the *most developed regions* were primarily able to apply for and accommodate the majority of funds available from the SAPARD programme successfully, our analysis confirms this development by showing a significant (at 0.05 level) positive correlation (0.43) between the intensity of programme support (measured *per region* basis) and the overall level of rural development measured in terms of the RDI (Table 18)[28].

27 These companies were also the most effective in submission of well-designed project proposals.

28 Also when calculating at per capita basis, the intensity of SAPARD funds was found to be significantly (at 0.05 level) and positively correlated with the overall level of rural development (yet, correlation was much lower =0.24).

Table 18: Slovakia: Correlation matrix between intensity of SAPARD (per region basis) and the RDI

	RDI 2002	SAPARD funds (total)
RDI 2002	1.0000	
SAPARD funds (total)	0.4303*	1.0000

9.4. Approaches for assessment of the impact of SAPARD programme

An analysis of the allocation of *total* SAPARD funds', in Slovakia shows that *all* NUTS-4 rural regions in the period 2002-2004 were, to some extent, supported (at least by one of 1-7 measures) from the SAPARD programme. However, the distribution of an individual (regional) intensity to programme exposure was highly skewed. In many cases the intensity of programme support (per region) was almost negligible (e.g. Poltar, Turcianskie Teplice, etc.). In 42% of Slovak regions the *total* programme support was lower than 66% of an average support measured per capita basis (i.e. lower than 600 SKK per capita, compared with 904 SKK per capita in regions' average).

With regard to the regional distribution of programme support linked to *individual programme measures* (1-7), the picture is slightly different. That is to say, in all examined cases (applies to each individual measure) the support from the SAPARD programme embraced only a *subset* of all NUTS-4 regions (i.e. in no single case did the programme support linked to a *specific* SAPARD measure embrace all Slovak regions). Additionally, many Slovak regions obtained the support from individual SAPARD measures that was below 66% of the country average (for a given measure).

Taking into consideration the above situation, the assessment of the impact of the SAPARD programme was carried out using two complementary approaches:

- Approach 1 (based on the *binary* PSM matching method) allowed the estimation of the effectiveness of the programme support by comparing regions that: a) received support from the programme with equivalent regions that did not receive any support from SAPARD, or b) received programme support above a certain threshold (e.g. above 66% of country's average) with those where programme intensity was much below the country average. This approach was applied basically to the assessment of the impact of individual programme's measures.

- Approach 2 (based on the application of the generalized propensity score matching and dose-response function), allowed the estimation of the impact of the total support from the SAPARD programme at various programme support levels. This approach was applied mainly to the assessment of the impact of total funds from the SAPARD programme (i.e. where all regions were supported).

9.5. Application of a binary PSM matching to the assessment of the impact of the SAPARD programme in Slovakia

9.5.1. Total SAPARD funds (all measures)

The application of Approach 1, including the setting of the threshold (66% of the country average per capita), resulted in the division of all NUTS-4 regions into two groups: a) 42 regions where support obtained from the SAPARD programme was above the threshold (600 SKK per capita), and b) 30 regions "non-SAPARD

supported" (with the level of programme support less than 600 SKK per capita)[29].

Comparison of these two groups of regions (applying 600 SKK per capita as a threshold) reveals significant differences in all major regional characteristics (factors 1-21) determining both the selection of individual regions into the programme as well as the effect of the SAPARD programme (Table 19).

The most obvious differences, except of the overall level of rural development (the RDI was much higher in the programme supported group D=1 compared with non-supported regions D=0), concern factors f4 (Agriculture and natural endowment, with a much higher intensity level in group D=1 compared with group D=0), f3 (Social conditions and living environment (incl. availability of dwelling; with a much higher level of endowment in group D=1 compared with D=0), f16 (Primary schools; with a much higher density level in group D=1 compared with D=0), and f1 (Spatial density of social and retail infrastructure (per km^2); with a much lower level in group D=1 compared with D=0).

In summary, the analysis shows that the huge majority of SAPARD funds were targeted to regions that as a whole were: a) strongly agriculture oriented, b) characterized by relatively good social conditions (including endowments with primary schools) and living environment (including dwellings), and c) exhibited a high level of rurality (i.e. lower spatial density of social and retail infrastructure) compared to regions with a low intensity of programme support.

Significant differences in socio-economic and environmental characteristics of programme supported and non-supported regions prove that any direct comparisons of selected impact indicators in regions supported by the programme with respective impact indicators in non-supported regions would result in a considerable selectivity bias and thus unreliable results.

The next step of the analysis aimed therefore at assessing the impact of the SAPARD programme by comparing the situation in regions supported by the SAPARD programme with a similar regions that were non-supported by the programme (thus enabling disentangling effects of the programme from other confounding factors).

This was done *separately for all individual SAPARD measures*. Firstly, appropriate (*measure specific*) control groups were selected (e.g. selecting non-supported regions that, in terms of their characteristics, were not statistically different from the group of supported regions). Secondly, by calculating ATT indicators and applying a conditional DID method (i.e. combining ATT with DID) to the assessment of SAPARD's impact on the overall level of rural development (measured in terms of the RDI) and rural unemployment. Thirdly, by computing ATE and ATU policy indicators showing the potential effectiveness of the extension of the SAPARD programme to other regions; and fourthly, by assessing the sensitivity of obtained results (impact of hidden bias).

29 In fact by dividing NUTS-4 regions into two groups ("supported" vs. "non-supported" regions) using above criterion we disregard the potential impact of very small SAPARD projects (i.e. below 333 thousand EUR per region).

Table 19: Slovakia: Differences between "supported" and "non-supported" regions (programme participation criterion: total SAPARD funds > 600 SKK per capita)

Variable	Mean		
	D=1 (42)	D=0 (30)	D(1) – D(0)
f1	-.1708475	.2350822	-0.405929
f2	.0306248	.1208217	-0.090196
f3	.2251119	-.378494	0.603605
f4	.3078212	-.361034	0.668855
f5	-.1071268	-.017086	-0.090040
f6	-.1089812	.1206531	-0.229634
f7	1.207971	1.091276	0.116695
f8	.1657196	.1415932	0.024126
f9	.086248	-.071788	0.158036
f10	-.0378856	.0925739	-0.130459
f11	-.0037055	-.168402	0.164696
f12	-.0677918	.1705341	-0.238325
f13	-.0806877	-.006927	-0.073760
f14	-.0472144	-.224209	0.176994
f15	-.0202304	.0006487	-0.020871
f16	.0705653	-.395640	0.466205
f17	-.0376514	-.175056	0.137404
f18	-.1143344	.1575503	-0.271884
f19	.0988766	-.188457	0.287333
f20	-.1107786	.0912989	-0.202077
f21	-.0830286	.131373	-0.21440
RDI (2002)	.0062703	-.124174	0.130444

9.5.2. Estimation of the propensity scores

Given the individual characteristics of NUTS-4 regions (factors f1-f21) and information about regions' participation in the SAPARD programme ("supported" and "non-supported" regions), the propensity scores (i.e. the conditional probability of a region's participation in the SAPARD programme) were estimated separately for all individual regions and individual measures using a logit function (1-0).

The results of the logit estimation (all SAPARD measures) for total SAPARD measures are shown in Table 20.

The above estimation results were used to calculate the individual propensity scores (the conditional probabilities of a region's participation in the SAPARD programme) for all 72 NUTS-4 regions.

Table 20: Slovakia: Results of logit estimation (all SAPARD measures; participation criteria: programme support above 600 SKK per capita)

sapardMall	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
f1	-.7540899	1.433468	-0.53	0.599	-3.563635	2.055455
f2	-.4388994	.4843735	-0.91	0.365	-1.388254	.5104552
f3	1.125374	.4312193	2.61	0.009	.2802001	1.970549
f4	1.444231	.5008184	2.88	0.004	.4626451	2.425817
f5	-.2464712	.3687212	-0.67	0.504	-.9691515	.4762091
f6	-.3418738	.4504012	-0.76	0.448	-1.224644	.5408963
f7	.3052294	1.74678	0.17	0.861	-3.118396	3.728855
f8	.2365941	.4647741	0.51	0.611	-.6743464	1.147535
f9	.6041546	.5171485	1.17	0.243	-.4094379	1.617747
f10	-.210518	.3748301	-0.56	0.574	-.9451714	.5241354
f11	.2522791	.6308867	0.40	0.689	-.984236	1.488794
f12	-.3747681	.3444701	-1.09	0.277	-1.049917	.300381
f13	-.4118278	.3880493	-1.06	0.289	-1.17239	.3487349
f14	.5624386	.4312731	1.30	0.192	-.2828413	1.407718
f15	-.2726117	.8090837	-0.34	0.736	-1.858387	1.313163
f16	1.106433	.5163714	2.14	0.032	.0943633	2.118502
f17	.2555963	.4973568	0.51	0.607	-.719205	1.230398
f18	-.5404006	.422999	-1.28	0.201	-1.369463	.2886623
f19	.8935932	.5412636	1.65	0.099	-.167264	1.95445
f20	-.3386763	.4950336	-0.68	0.494	-1.308924	.6315718
f21	-.3951575	.3735557	-1.06	0.290	-1.127313	.3369981
_cons	.5387786	2.195699	0.25	0.806	-3.764713	4.84227

Logistic regression	Number of obs	LR chi2(21)	Prob > chi2	Log likelihood	Pseudo R2
	= 72	= 41.71	= 0.0046	= -28.045468	= 0.4265

9.5.3. Selection of matching algorithms and testing balancing property

Given the considerable differences between individual characteristics (factors f1-21) in supported and non-supported groups of regions the binary PSM matching was applied in order to find appropriate controls.

The binary PSM method balances the observed covariates between the supported group and a control group based on the similarity of their predicted probabilities of receiving support (e.g. above the threshold) from the SAPARD programme. Implementing common support conditions ensures that any combination of characteristics observed in the treatment group can also be observed among the control group. In our study a common support region was imposed on both sides, i.e. by dropping treatment observations whose estimated propensity scores is higher than the maximum or lower than the minimum propensity score of the controls and vice versa (i.e. dropping control observations whose estimated propensity score is higher than maximum or lower than minimum propensity score of the treated)[30]. In case areas of common

30 This was necessary in order to estimate both ATT as well as ATE.

Table 21: Slovakia: Division of regions after imposing common support conditions

Treatment assignment	Common support		
	off support	on support	total
Treated	15	15	30
Untreated	20	22	42
Total	35	37	72

support were not found (the support of X did not overlap for the participants and non-participants), respective NUTS-4 regions **i** were sorted out and matching was performed over the region of common support only.

For programme participation measured on a per capita basis (all SAPARD measures above a threshold 600 SKK per capita), imposition of the common support condition resulted in disregarding 35 "non-comparable" regions[31] (i.e. 15 non-supported regions and 20 programme supported regions), a selection of *comparable* 22 regions supported by the SAPARD programme, and 15 control regions (Table 21).

As the probability of observing two units with exactly the same value of propensity score is, in principle, zero, an estimation of programme effects requires using appropriate matching algorithms. The latter define the measure of proximity in order to define programme non-participants who are acceptably close (e.g. in terms of the propensity score) to any given programme participant. Given that the choice of both matching method (e.g. nearest neighbour (NN) matching, calliper matching, Gaussian kernel matching, Epanechnikov matching, etc) and selection of an appropriate matching parameter (e.g. number of nearest neighbours in NN matching, radius size in calliper matching, bandwidth size in Gaussian or Epanechnikov matching, etc.) can make a difference in small samples, and the quality of a given matching technique depends strongly on a dataset, the selection of a relevant matching technique in our study was carried out using the following three criteria: i) standardized bias (Rosenbaum and Rubin, 1985); ii) t–test (Rosenbaum and Rubin, 1985); and iii) joint significance and pseudo R^2 (Sianesi, 2004).

Given the above criteria, the best results concerning selection of an appropriate matching algorithm were achieved by applying a two-step selection procedure. Firstly, by scaling respective matching parameters *within each matching algorithm* (e.g. the number of neighbours in the nearest neighbour algorithm; size of calliper in calliper matching; size of bandwidth in kernel Gaussian; size of bandwidth in kernel Epanechnikov, etc.) and applying a linear search to find those matching parameters under each matching algorithm that minimize the estimated standardized bias[32] (after matching). Secondly, by searching across all considered matching algorithms and applying the min{min} criterion as the main final selection option.

In all cases (i.e. various matching algorithms)[33] an optimal solution could easily be found due to local/global convexity of the objective function with respect to adjusted matching parameters (e.g. radius magnitude

31 Outside of the imposed common support area

32 The standardized bias is the difference of the sample means in the treated and non-treated (full or matched) sub-samples as a percentage of the square root of the average of the sample variances in the treated and non-treated groups (Rosenbaum and Rubin, 1985).

33 This does not apply to local linear weighting function matching which first smoothes the outcome and then performs nearest neighbor matching. In this case more controls are used to calculate the counterfactual outcome than the nearest neighbor only (Leuven and Sianesi, 2007).

Table 22: Slovakia: Comparison of matching algorithms (participation criterion: support per capita; impact indicator: RDI in 2002)

Matching method	Matching parameters	Estimated standardized bias (after matching)
Nearest neighbours	N (1)	16.401
	N (2)	12.508 → min
	N (2)	12.94
Radius caliper	(0.24)	11.161
	(0.25)	11.156 → min
	(0.26)	11.245
Kernel normal (Gaussian)	bandwidth (0.27)	10.788
	bandwidth (0.28)	10.781 → Selection Min {Min}
	bandwidth (0.29)	10.791
Kernel biweight		13.888
Kernel epanechnikov	bandwidth (0.34)	11.055
	bandwidth (0.35)	11.052 → min
	bandwidth (0.36)	11.064

in radius matching; or the number of nearest neighbours in nearest neighbour matching, etc.). An overview of results from the selection procedure involving various matching algorithms is provided in Table 22.

By applying the above selection procedure to our data (conditional regional participation in the SAPARD programme given covariates f1-f21) we found that a kernel matching (Gaussian bandwidth (0.28)) was that one that ensured the minimization of the standardized selection bias (after matching) and thus the highest reduction of selection bias, and at the same time satisfaction of both the balancing property test (t-test) as well as pseudo R^2 tests (see Tables 23 and 24).

The balancing property test shows that, compared with the situation prior to the matching, application of the above matching procedure led to the selection of an appropriate control group of regions (performed t-tests confirmed the elimination of all significant differences between individual regional characteristics in both groups of regions and therefore significant reduction of the selection bias). This applies both to the differences in the RDI and all important variables (factors) determining both programme participation and programme outcomes, e.g. F4 (Agriculture and natural endowment), F1 (Spatial density of social and retail infrastructure (per km^2), F6 (Spatial density of public utilities and social infrastructure, gas pipelines, water-supply-system (per km^2), F9 (Hotels and recreation facilities), etc.

Also other tests, e.g. pseudo R^2 (pseudo R^2 = 0.43 before matching and pseudo R^2= 0.23 after matching) fully confirmed the applicability of the above approach (Table 24).

Table 23: Slovakia: Covariates' balancing test between selected (common support region; kernel Gaussian matching bw 0.28) programme supported and non-supported NUTS-4 regions (programme intensity per region basis)

		Mean		%reduct		t-test	
Variable	**Sample**	**Treated**	**Control**	**%bias**	**bias**	**t**	**p> t**
f1	Unmatched	-.17085	.23508	-37.9		-1.74	0.087
	Matched	-.19018	-.19523	0.5	98.8	0.06	0.955
f2	Unmatched	.03062	.12082	-8.9		-0.37	0.716
	Matched	.02569	-.21683	23.9	-168.9	0.77	0.449
f3	Unmatched	.22511	-.37849	60.4		2.63	0.011
	Matched	-.07682	-.01097	-6.6	89.1	-0.27	0.789
f4	Unmatched	.30782	-.36103	71.5		2.89	0.005
	Matched	-.07531	-.43527	38.5	46.2	0.96	0.344
f5	Unmatched	-.10713	-.01709	-9.8		-0.39	0.696
	Matched	.10174	-.06301	18.0	-83.0	0.41	0.681
f6	Unmatched	-.10898	.12065	-22.7		-0.96	0.342
	Matched	-.04513	-.1159	7.0	69.2	0.30	0.764
f7	Unmatched	1.208	1.0913	32.6		1.33	0.189
	Matched	1.1958	1.1358	16.8	48.5	0.38	0.706
f8	Unmatched	.16572	.14159	2.5		0.10	0.917
	Matched	.03939	.26792	-23.5	-847.2	-0.53	0.602
f9	Unmatched	.08625	-.07179	16.0		0.65	0.521
	Matched	-.08267	-.17201	9.0	43.5	0.41	0.687
f10	Unmatched	-.03789	.09257	-12.2		-0.53	0.600
	Matched	.06903	.02863	3.8	69.0	0.11	0.914
f11	Unmatched	-.00371	-.1684	16.6		0.66	0.511
	Matched	-.13596	-.04801	-8.9	46.6	-0.28	0.782
f12	Unmatched	-.06779	.17053	-22.9		-0.97	0.334
	Matched	.09025	.2778	-18.0	21.3	-0.48	0.635
f13	Unmatched	-.08069	-.00693	-7.5		-0.32	0.750
	Matched	-.10433	-.09691	-0.8	89.9	-0.06	0.951
f14	Unmatched	-.04721	-.22421	19.5		0.81	0.421
	Matched	-.12468	-.13783	1.4	92.6	0.09	0.932
f15	Unmatched	-.02023	.00065	-2.0		-0.09	0.927
	Matched	-.01171	.06336	-7.3	-259.5	-0.37	0.711
f16	Unmatched	.07057	-.39564	46.8		1.98	0.052
	Matched	-.16771	-.15401	-1.4	97.1	-0.09	0.928
f17	Unmatched	-.03765	-.17506	14.1		0.60	0.553
	Matched	-.22372	-.13078	-9.5	32.4	-0.49	0.624
f18	Unmatched	.11433	.15755	-29.0		-1.23	0.223
	Matched	-.10495	-.04217	-6.7	76.9	-0.05	0.960
f19	Unmatched	.09888	-.18846	32.1		1.35	0.180
	Matched	.00188	.06309	-6.8	78.7	-0.19	0.848
f20	Unmatched	-.11078	.0913	-19.7		-0.80	0.425
	Matched	.13729	-.05072	18.4	7.0	0.77	0.445
f21	Unmatched	-.08303	.13137	-21.4		-0.92	0.360
	Matched	-.19943	-.24333	4.4	79.5	0.10	0.918
RDI2002	Unmatched	.00627	-.12417	52.2		2.14	0.036
	Matched	-.08829	-.10414	6.3	87.9	0.05	0.963

Table 24: Slovakia: Results of pseudo R^2 tests

Sample	Pseudo R2	LR chi2	p>chi2
Unmatched	0.431	42.20	0.004
Matched	0.229	11.43	0.954

9.5.4. Calculation of policy evaluation parameters (ATT, ATE, ATU)

A comprehensive assessment of a programme's impact requires separation of various programme effects of which the most important are: a) effect on regions which participated in a given programme (Average Treatment Effect on the Treated - ATT); b) effect on an average region randomly selected from the pool of programme participants and non-participants (Average Treatment Effect – ATE) and c) effect of the programme on the regions which did not participate (Average Treatment Effect on the Untreated – ATU).

In our study, the above policy evaluation parameters (ATT, ATE, and ATU) were calculated on the basis of the estimated propensity scores using the following impact indicators:

a. The RDI
b. Unemployment (absolute values)
c. Unemployment (per capita)

The results of ATT, ATE and ATU calculations are shown in Table 25. Given these parameters the programme impact was quantified using a *conditional DID estimator*, i.e. combining PSM (ATT, ATE, and ATU) and difference in differences (DID) methods.

9.5.5. Conditional DID estimator

Application of the conditional DID estimator to the assessment of the programme impact at the regional level shows that the overall impact of the SAPARD programme in Slovakia on the level of regional development, as well as on rural unemployment, were negligible.

In fact, our results show that in regions that obtained low support from SAPARD (i.e.

Table 25: Slovakia: Estimated policy evaluation parameters (per capita basis)

Calculation basis	RDI			Unemployment (absolute)			Unemployment (per capita)		
	2002	2005	DID (2005 - 2002)	2002	2005	DID (2005 - 2002)	2002	2005	DID (2005 - 2002)
Unmatched 1 ( )	.006270	.0910252	.00847552	7136	4664	-2472	.100763	.068188	-.032575
Unmatched 0 ()	-.12417	-.020165	.104005	6806	4587	-2219	.101956	.070347	-.031609
Difference (1-0)	.130444	.1111904	-.019254	329	76	-253	-.00119	-.00215	-.00096
Matched M 1 ()	-.088289	-.001828	.086461	6889	4890	-1999	.10065	.071229	-.029421
Matched M 0 ()	-.104754	.0306261	.1353801	6406	4279	-2127	.103300	.069738	-.033562
ATT	.016464	-.032455	-.048919	483	610	127	-.00264	.001490	.00413
ATU	-.005894	-.057750	-.051856	507	582	75	-.00092	.002313	.003233
ATE	.007400	-.042709	-.0501093	493	599	106	-.00194	.001824	.003764

below a 600 SKK per capita from all SAPARD measures) improvement of the overall level of rural development (the RDI) and unemployment indicators were generally faster than in comparable regions which received the highest programme support (above 600 SKK per capita). This means that the impact of the SAPARD programme on a general performance (overall level of rural development and unemployment) in well-developed Slovak regions was negligible.

9.6. Impact of SAPARD programme (by measures)

A slightly differentiated picture concerning the effectiveness of the SAPARD programme in Slovakia was obtained by carrying out an estimation of the programme's impact at *individual measures basis* (Table 26).

Our results show that out of 1-7 measures examined, only two individual SAPARD measures (i.e. Measure 1: investment in agricultural enterprises, and Measure 6: Agricultural production methods designed to protect the environment and maintain the countryside) contributed positively to the overall level of rural development in supported regions (measured in terms of the RDI). On the other hand, the implementation of the measure M5 (Forestry) was found to be highly ineffective (the RDI was negatively affected)[34].

Table 26: Slovakia: Estimated impact of SAPARD (by measures) using a binary PSM method

Measure	Overall growth (RDI)			Unemployment (absolute number)			Unemployment (per capita)		
	ATT (2002)	ATT (2005)	Impact (Cond. DID)	ATT (2002)	ATT (2005)	Impact (Cond. DID)	ATT (2002)	ATT (2005)	Impact (Cond. DID)
M 1	.05100	.07389	+	-928.7	-399.7	- -	-.0095	-.00370	-
M 2	.04130	.03457	-	492.7	601.3	-	-.0084	-.00237	-
M 4a	.00341	-.0086	-	3444	1679	+++	.00672	-.00080	+
M 4b	.06113	.03813	-	595.2	304.5	+	.00206	.001736	+
M 5	.00304	-.04819	--	2015	1601	+	.008416	.008238	++
M 6*	.10492	.18014	+++	-3965	-2358	- - -	-.03151	-.02211	-
M 7	-.00333	-.0073	-	1417	753.8	++	.001866	-.00178	+

Measures: M1: Investment in agricultural enterprises; M2: Improving the processing and marketing of agricultural and fishery products; M4a: Investments not involving infrastructure; M4b: Investments in infrastructure not bringing substantial revenues; M5: Forestry; M6: Agricultural production methods designed to protect environment and maintain the countryside; M 7: Land improvement and reparcelling.

** Pseudo R test rejected (small number of observations)*

Thresholds: M1: D=1 if M1 > 2 Mill SKK per region, D=0 otherwise; M2: D=1 if M2 > 4 Mill SKK per region, D=0 otherwise; M4a: D=1 if M4a > 0, D=0 otherwise; M4b: D=1 if M4b > 4 Mill SKK per region, D=0 otherwise; M5: D=1 if M5 > 0, D=0 otherwise; M6: D=1 if M6 > 0, D=0 otherwise; M7: D01 if M7 > 5 Mill SKK per region, D=0 otherwise.

34 Original funds allocation to Measure 5 (forestry) was several times higher than at the end of the SAPARD programme. Out of 35 contracted projects in the forestry sector two major projects (approximately 16 Mill SKK) were suspended due to bankruptcy of contracted forest enterprises. Average amount per project under Measure 5 was the lowest from all average project costs under other measures. No result indicators under Measure 5 set in the RDP plan were monitored. No measure 5 impact indicators were set and monitored. See: Ex-post evaluation of the SAPARD programme in the Slovak Republic. P.C.M. Group, December 2007.

In terms of the impact of SAPARD measures on rural unemployment, Measures 4a, 4b, 5, and 7 were found to have a positive impact on the reduction of rural unemployment (measured both in absolute terms and per capita basis). Measure 4a had an especially positive impact on the reduction of rural unemployment (Investments not involving infrastructure) that was mainly focused on support of local agro-tourist facilities. On the other hand, due to the introduction of technological advancements, implementation of SAPARD measures M1 (investment in agricultural enterprises), M2 (investment in food processing) and M6 (environmental investments) had a negative impact on unemployment, i.e. the above measures were found to lead to an increase of rural unemployment.

9.7. Assessment of the impact of the SAPARD programme using a generalized propensity score and dose-response function approach

The application of a generalized propensity score matching and dose-response function approach is particularly advantageous if the huge majority of regions, or all regions, are subject to support from the programme (low number or no D=0). Additionally, the GPS approach allows questions relating to marginal programme effects to be answered (by linking programme impacts to the level of programme intensity).

Application of the GPSM methodology to the analysis of the impact of the SAPARD programme in Slovakia was carried out using information about programme intensity on a *per region* and *per capita* basis. The four main steps were: a) estimation of the treatment function; b) calculation of the GPS and carrying out balancing tests; c) modelling conditional expectations of the programme outcome; and d) calculation of the dose-response and derivative dose-response functions.

9.7.1. Estimation of the treatment function

Given regional individual covariates (f1-f21) and the regional levels of programme intensity (per capita) the conditional treatment function was estimated according to a modified eq. 14

$$Ln\_t | X_i \approx N(\beta_0 + \beta_1' X_i, \sigma^2)$$

Where: X = covariates (f1-f21)

Ln_t = logarithm of programme intensity per capita level

Results of the estimation of the conditional treatment function are shown in Tables 26a and 26b.

Table 26a: Slovakia: Results of estimated conditional treatment function (programme intensity measured per capita basis)

ln_sapardn	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
f1	-.168177	.0999547	-1.68	0.099	-.368942	.0325879
f2	-.2433884	.0988791	-2.46	0.017	-.4419929	-.0447839
f3	.1017886	.097278	1.05	0.300	-.0936001	.2971773
f4	.4000884	.092825	4.31	0.000	.2136439	.586533
f5	.0124345	.111807	0.11	0.912	-.2121364	.2370055
f6	.0319413	.0963422	0.33	0.742	-.1615677	.2254502
f7	.3659093	.4572321	0.80	0.427	-.5524685	1.284287
f8	-.056359	.1304051	-0.43	0.667	-.3182853	.2055674
f9	.0373763	.0926989	0.40	0.689	-.148815	.2235675

f10	-.0132258	.0923704	-0.14	0.887	-.1987572	.1723057
f11	.0490882	.101816	0.48	0.632	-.1554153	.2535917
f12	.0109451	.094692	0.12	0.908	-.1792494	.2011396
f13	-.1671369	.1000208	-1.67	0.101	-.3680346	.0337609
f14	.0693514	.1103534	0.63	0.533	-.1522999	.2910027
f15	.1769276	.1044731	1.69	0.097	-.0329127	.3867679
f16	.3256221	.103869	3.13	0.003	.1169951	.5342491
f17	.148131	.1204114	1.23	0.224	-.0937224	.3899844
f18	-.0902968	.1079426	-0.84	0.407	-.3071059	.1265122
f19	.3246935	.1091284	2.98	0.004	.1055026	.5438844
f20	-.089376	.0897081	-1.00	0.324	-.2695599	.090808
f21	-.3169725	.0968343	-3.27	0.002	-.5114699	-.1224751
_cons	6.102766	.5727254	10.66	0.000	4.952413	7.253119

Table 26b: Slovakia: Supplementary information on results of estimated conditional treatment function (programme intensity measured per capita basis)

Estimation of the propensity score				
Source	SS	df	MS	Number of obs = 72
Model	47.5510971	21	2.26433796	F(21, 50) = 3.70
Residual	30.6290118	50	.612580237	Prob > F = 0.0001
				R-squared = 0.6082
Total	78.1801089	71	1.10112829	Adj R-squared = 0.4437
				Root MSE = .78268

9.7.2. Calculation of GPS and balancing property tests

Obtained estimates (9.7.1.) were used to calculate region specific propensity scores (prior to the programme) according to eq 15. Testing of the balancing properties for covariates was performed using a method proposed in Hirano and Imbens (2004), i.e. by blocking on both the treatment variables (e.g. programme intensity per capita) and on the estimated GPS. Given GPS and various intensity levels of the SAPARD programme support per region (on per capita basis), the balancing property test (t-test) was carried out for all variables f1-f21 in pre-specified blocks of GPS (=2) and programme intensity levels (=3), i.e. by testing if for each GPS block the covariate means of regions belonging to the group of the particular intensity level of programme support are significantly different from those of regions with a different intensity level of support, but similar GPS level. The results of the t-tests developed in Bia and Mattei (2007) showed that balancing property was satisfied for all variables, GPS blocks and intensity levels.

9.7.3. Modelling the conditional expectation of the programme outcome and dose-response function

Given Ti and estimated GPS (Ri) for each of the NUTS-4 regions in Slovakia, the conditional expectation of the programme outcome measured in terms of the RDI (Y = Δ RDI) was modelled as a flexible function of its two arguments (Ti and Ri) according to eq (18) (polynomial quadratic function).

Results of estimated the conditional expectation of the outcome function <E[Y(t)]> are shown in Table 27.

While the estimated coefficients in this model do not have a causal interpretation (Hirano and

Table 27: Slovakia: Estimated parameters of the conditional expectation of the outcome function

Δ RDI (2002-2005)	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
/b0	.0654284	.0433889	1.51	0.136	-.0212004	.1520572
/b1	-.0000474	.000054	-0.88	0.383	-.0001551	.0000603
/b2	-.1084254	.0576379	-1.88	0.064	-.2235031	.0066524
/b3	1.79e-08	2.15e-08	0.83	0.407	-2.50e-08	6.08e-08
/b4	-.0309421	.0142702	-2.17	0.034	-.0594334	-.0024508
/b5	.0000147	.0000209	0.70	0.485	-.0000271	.0000565

Table 28: Slovakia: Estimated dose-response function and the derivative dose response function for SAPARD programme. Impact indicators: change in the RDI; change in unemployment. (all measures; programme intensity on per capita basis)

Level of support SKK/ capita	RDI				Unemployment (absolute)			
	Dose-response function E[Y(t)]	E[Y(t+1)]	E[Y(t+1)] - E[Y(t)]	Derivative Dose-response E[Y(t+1) - Y(t)]	Dose-response function E[Y(t)]	E[Y(t+1)]	E[Y(t+1)] - E[Y(t)]	Derivative Dose-response E[Y(t+1) - Y(t)]
2702	-0.620799	-0.6208121	-0.0000123	-0.0000123	1366	1366	.298584	.2986673
2562	-0.565194	-0.5652093	-0.0000146	-0.0000146	1029	1029	.2301025	.2300991
2423	-0.510655	-0.5106728	-0.0000169	-0.000017	706	706	.1624756	.1624718
2283	-0.457296	-0.4573159	-0.0000193	-0.0000193	396	396	.0957031	.095687
…								
1572	-0.209013	-0.2090443	-0.0000305	-0.0000305	-947	-947	-.233886	-.233878
1552	-0.202534	-0.2025654	-0.0000308	-0.0000308	-980	-980	-.243286	-.243340
1470	-0.177078	-0.1771105	-0.0000321	-0.0000321	-1012	-1012	-.252807	-.252796
1449	-0.170832	-0.1708648	-0.0000324	-0.0000324	-1044	-1044	-.262451	-.262338
…								
998	-0.046814	-0.0468547	-0.0000403	-0.0000402	-1706	-1707	-.515380	-.515387
977	-0.041855	-0.0418963	-0.0000407	-0.0000407	-1727	-1727	-.526855	-.526825
895	-0.022706	-0.0227486	-0.0000424	-0.0000424	-1746	-1747	-.538452	-.538465
875	-0.018099	-0.018142	-0.0000429	-0.0000429	-1766	-1766	-.550293	-.550323
…								
772	0.0037266	0.0036813	-0.0000454	-0.0000454	-1897	-1897	-.655273	-.655268
402	0.0483346	0.0482733	-0.0000613	-0.0000613	-1825	-1826	-1.05456	-1.05452
300	0.0243507	0.0242796	-0.000071	-0.000071	-1422	-1423	-1.28442	-1.28445
279	0.0125944	0.0125208	-0.0000737	-0.0000737	-1270	-1271	-1.34606	-1.34609
…								
259	-0.003244	-0.0033215	-0.0000766	-0.0000766	-1076	-1078	-1.41540	-1.41537
95	-0.736462	-0.7365928	-0.0001302	-0.0001301	6358	6355	-2.64111	-2.64099
65	-1.46721	-1.467369	-0.0001587	-0.0001586	13374	13371	-3.28613	-3.28659

Imbens, 2004), the estimated regression function is later used to estimate of the causal effects of the SAPARD programme (average programme effects and marginal outcome of programme support at particular level T).

The dose-response function (DRF) was computed as the *average* potential outcome for *each level of treatment* according to eq 16. The marginal programme effects were estimated means a *derivative dose-response function* E[Y(t+1) - Y(t)]. The bootstrap methods were applied to obtain standard errors that take into account the estimation of GPS and the parameters of the estimated conditional expectation of the outcome function. Results of the estimated dose-response and derivative dose-response function are shown in Table 28.

Application of the GPS and the dose-response function to the assessment of the impact of the SAPARD programme in Slovakia primarily confirms the results obtained by using the binary PSM method, i.e. it proves that the impact of SAPARD measures (total funds) on the overall level of rural development (measured in terms of the RDI) across Slovak regions was generally negligible (or negative), except for those regions which received programme support between 260-780 SKK per capita (positive dose-response function). Apparently, the positive impact on the overall level of rural development (measured in terms of the RDI) of two SAPARD measures M6 (Agricultural production methods designed to protect environment and maintain the countryside) and M1 (modernization of agricultural enterprises), could not overcompensate some negative effects stemming from implementation of other SAPARD measures, especially M5 (forestry).

More positive impacts of all SAPARD measures were found on a reduction of rural unemployment. Obviously, *reduction* of the number of unemployed caused by measures M4a (Investments not involving infrastructure, mainly in agro-tourism), M7 (Land improvement and re-parcelling), M4b (Investments in rural infrastructure not bringing substantial revenues) and M5 (forestry) overcompensated an *increase* of unemployment caused by measures: M1 (Investment in agricultural enterprises), M2 (Improving the processing and marketing of agricultural and fishery products, mainly investment in food industry), and M6 (Agricultural production methods designed to protect environment and maintain the countryside).

Our results show that in those regions that received programme support between 259-1573 SKK per capita, the impact of the SAPARD programme on rural unemployment was positive (i.e. SAPARD funds contributed to a reduction in the number of unemployed persons). While the highest reduction of rural unemployment was found in regions with programme intensity in the range between 402-998 SKK per capita, the effectiveness of the programme intensity above 2280 SKK per capita (the highest support level) and those below 90 SKK per capita (i.e. the lowest support level) was found to be negative (i.e. in those regions the SAPARD programme contributed to an increase in the number of unemployment persons).

10. Conclusions

The basic objective of this study was to analyze the impact of EU RD programmes on rural regions. Aggregated effects of a given RD programme at regional levels were estimated using the Rural Development Index (RDI) – a proxy describing the overall quality of life in individual rural areas. The weights of economic, social and environmental domains entering the RDI index (composite indicator) were derived empirically from the econometrically estimated intra- and inter-regional migration function after selecting the "best" model from various alternative model specifications (e.g. panel estimate logistic regression nested error structure model, spatial effect models, etc). The impacts of individual RD measures were analysed by means of a counterfactual analysis by applying combination of the binary Propensity Score Matching (PSM) (e.g. Kernel matching) and difference-in-differences (DID) methods (i.e. by comparing supported regions and matched control group, prior to the programme and after it). Evaluation of programme effects (by programme measures) at regional level is carried out on the basis of the estimated policy parameters: Average Treatment Effects (ATE), Average Treatment on Treated (ATT) and Average Treatment on Untreated (ATU) effects by using the RDI Index and unemployment ratios as impact indicators. Given information on regional intensity to programme exposure (financial input flows by regions) the *overall* impact of obtained support via a given RD programme was estimated by means of a dose-response function and derivative dose-response function within the framework of a generalized propensity score matching (GPS). Furthermore, sensitivity analysis (Rosenbaum bounds) was carried out in order to assess a possible influence of unobservables on obtained results (under a binary PSM methodology). Above methodologies were empirically applied to evaluation of the impact of the SAPARD programme in Poland and Slovakia in years 2002-2005 at NUTS-4 level.

Our results show that the application of the GPS and the dose response function to the assessment of the impact of a given RD programme using the RDI combined with other partial indicators as an impact measure enables a more precise estimation of the effects of the given programme, compared with traditional "naive" evaluation techniques or methodologies based on binary PSM methods.

The major advantages from applying the RDI as an impact indicator in the framework of a generalized propensity score approach to the evaluation of RD programmes are as follows:

- The approach allows for considering of all potential effects of a given RD programme (aggregated or separated by programme measures) on various rural development domains (economic, social, environmental, etc.) and on the overall quality of life of population living in individual rural areas.

- The approach incorporates (implicitly) numerous general equilibrium effects of a programme, e.g. multiplier effects, substitution effects, into the analysis.

- While the weights applied into the construction of the RDI represent society's valuation of endowments and socio-economic trends observable at local/regional levels (estimated weights are representative for society as whole i.e. reflect both the decision of the migrating population and of the population that stays in the region) an application of the above weighting system allows for a more comprehensive assessment

of *social costs and benefits* of a given programme.

- The GPS is especially applicable in cases, when the probability of receiving a given level (intensity) of support is expected to depend on the intensity/distribution of individual regions' characteristics.

- The GPS extends and improves the quality of the analysis of programme effects compared to a binary PSM-DID method. Especially promising is the possibility of the estimation of the average and marginal potential outcomes that correspond to specific values of continuous programme doses (i.e. for each level of programme support) by means of a dose-response and derivative dose-response functions. Here, programme impacts are linked to the level of programme intensity.

- An essential advantage of the proposed methodology is that GPS method eliminates (or at least substantially reduces) selection bias and allows to estimate individual programme effects not only in "average" terms, but also for different programme support intensity levels (!).

- The above evaluation methodology permits testing a number of common stipulations, e.g. positive effect of a given policy on various indicators of regional performance, e.g. employment, labour productivity, environmental and social indicators, etc.

- The major weakness is that the above method requires an abundant and good quality data (available at regional levels) and considerable technical skills on side of its users (e.g. programme evaluators).

Clearly, the above methodology is highly applicable both for analysis of effects of RD as well as structural programmes at a regional level, and is powerful both at the aggregated level (e.g. NUTS 2) as well as NUTS 3 or NUTS 4 levels.

11. References

Aakvik, A. " Bounding a Matching Estimator: The Case of a Norwegian Training Program," *Oxford Bulletin of Economics and Statistics*, Department of Economics, University of Oxford, 2001, 63(1), pp. 115-43.

Abadie, A. and Imbens, G. "Large sample properties of matching estimators for average treatment effects." *Working paper*, 283. http://elsa.berkeley.edu/ ,2004.

Baker, J. H. "Evaluating the impact of development projects on poverty. A Handbook for practitioners", World Bank, 2000.

Becker, S. and Caliendo, M. "mhbounds - Sensitivity Analysis for Average Treatment Effects," Discussion Papers of DIW Berlin 659, DIW Berlin, German Institute for Economic Research, 2007.

Bia, M. and Mattei, A. "Application of the Generalized Propensity Score. Evaluation of public contributions to Piedmont enterprises," *P.O.L.I.S. department's Working Papers 80*, Department of Public Policy and Public Choice - POLIS, 2007.

Bondonio, D. "Evaluating decentralized policies: a method to compare the performance of economic development programmes across different regions or states," *Evaluation ,* 2002, 8(1), pp. 101-124.

Bourguignon, F. and Silva, L. A. P. d. "Evaluating the poverty and distributional impact of economic policies: a compendium of existing techniques," In: World Bank (eds.), *The impact of economic policies on poverty and income distribution.* World Bank [u.a.]: Washington, DC [u.a.], pp. 1-24, 2003.

Bourguignon, F. and Ferreira, F. H. G. "Ex ante evaluation of policy reforms using behavioral models," In: World Bank (eds.), *The impact of economic policies on poverty and income distribution*. World Bank [u.a.]: Washington, DC [u.a.], pp. 123-141, 2003.

Bryson, A., "The Union Membership Wage Premium: An Analysis Using Propensity Score Matching", Centre for Economic Performance, LSE, *CEP Discussion Papers 0530*, 2002.

Bryson, A. et.al, "The use of propensity score matching in the evaluation of active labour market policies.", Policy Studies Institute, U.K. Department for Work and Pensions Working *Paper No. 4*. http://www.dwp.gov.uk/asd/asd5/wp-index.html, 2002.

Burtless, G., "International Trade and the Rise in Earnings Inequality", Journal of Economic Literature, 1995, 33, pp. 800-816.

Caliendo, M. and Kopeinig, S. "Some Practical Guidance for the Implementation of Propensity Score Matching.", *Discussion Paper 485,* DIW Berlin, IZA Bonn, 2005.

Caliendo, M. et. al. "The Employment Effects of Job Creation Schemes in Germany: A Micro econometric Evaluation," *IZA Discussion Papers 1512*, Institute for the Study of Labour (IZA), 2005.

Caliendo, M. et. al. "Identifying Effect Heterogeneity to Improve the Effiency of Job Creation Schemes in Germany?," *ZEW Discussion Papers 05-21*, ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research, 2005.

CEAS (2003), Ex-post evaluation of measures under regulation (EC) No 951/97 on improving the processing and marketing conditions of agricultural products, Agra CEAS Consulting, September.

Cochran, W. and Rubin, D. "Controlling Bias in Observational Studies: A Review" Sankhya,1973, 35, pp. 417-46.

Dehejia, R. H. and Wahba, S. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association,* 1999, 94, pp. 1053–1062.

Dehejia, R. H. and Wahba, S. "Propensity Score-Matching Methods For Nonexperimental Causal Studies," *The Review of Economics and Statistics,* 2002, 84, pp. 151-161.

Del Corpo, B., U. Gasparino, E. Bellini, W. Malizia, (2008), "Effects of Tourism Upon the Economy of Small and Medium-Sized European Cities: Cultural Tourists and 'The others'", *FEEM Working Paper No. 44.2008, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1140611*

DiPrete, T. and M. Gangl, M. "Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments." *Sociological methodology*, 2004, 34, pp.271 – 310.

European Commission (2002a), Guidelines for the mid term evaluation of rural development programmes 2000-2006 supported from the European Agricultural Guidance and Guarantee Fund. DOC. STAR VI/43517/02.

European Commission (2002b), Guidelines for the mid-term evaluation of rural development programmes funded by SAPARD 2000-2006, DG AGRI.

European Commission, EC (2002d), Impact Assessment of the Mid-term Review Proposals for Agricultural Markets and Revenues in the EU-15 and EU-25 using the ESIM Model, DG AGRI, December.

European Commission, EC (2011), Monitoring and Evaluation for CAP post 2013. Background document for stakeholder conference held in Brussels on 20-21 September 2011.

Ederveen, S.; Gorter, J.; Mooij, R. A. d. and Nahuis, R. *Funds and games: the economics of European cohesion policy.* Occasional paper, No. 3, ENEPRI, 2003.

Essama-Nssah, Boniface. "Propensity score matching and policy impact analysis: a demonstration in EViews." Washington, DC: World Bank, Poverty Reduction and Economic Management Network, Poverty Reduction Group, 2006.

EuroCARE (2002a), Impact analysis of the European Commission's Proposal under the Mid-term Review of the CAP using the CAPSIM model.

EuroCARE (2002b), Mid–term review proposal impact analysis with the CAPRI modelling system, Department of Economics and Agricultural Policy, University of Bonn.

FAPRI (2002), FAPRI Analysis of the European Commission's Mid-term Review Proposals, Food and Agricultural Policy Research Institute, University of Missouri, December.

Felici, F., Paniccia, R. and Rocchi, B., "Economic Impact of Rural Development Plan 2007 2013 in Tuscany", 2008 International Congress, August 26-29, 2008, Ghent, Belgium 44256, European Association of Agricultural Economists.

Froelich, M. "A Note on the Role of the Propensity Score for Estimating Average Treatment Effects," *Econometric Reviews*, 2004, 23(2), pp. 167-174.

FRÖLICH (2004b): Finite sample properties of propensity-score matching and weighting estimators, Review of Economics and Statistics, 86(1), 77-90)

Goldstein, M, "An introduction to impact evaluation", *www1.worldbank.org/publicsector/decentralization/.../**Goldstein**.ppt,* 2007

Heckman, James J. and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica,* 2005, 73 (3), pp. 669–738.

Heckman, J.," Randomization as an Instrumental Variable Estimator", *Review of Economics and Statistics*, 1996, 56, pp. 336-341.

Heckman, James J., Ichimura, Hidehiko and Todd, Petra E. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," Review of Economic Studies, Blackwell Publishing, 1997, 64(4), pp. 605-54.

Heckman, James, Ichimura, Hidehiko, Smith, Jeffrey and Todd, Petra "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, 66 (5), pp. 1017–1098.

Hirano, K. and Imbens, G., (2004) The Propensity score with continuous treatment, *chapter for Missing data and Bayesian Method in Practice*: Contributions by Donald Rubin Statistical Family.

Holland, Paul. W. "Statistics and Causal inference (with discussion"*), Journal of the American Statistical Association,* 1986, 81, pp. 945-970.

Imai, K. and van Dyk, D.A. "Causal inference with general treatment regimes: Generalizing the propensity score", *Journal of the American Statistical Association*, 2004, 99, pp. 854-866.

Imbens, Guido and Wooldridge, Jeffrey. "Recent Developments in the Econometrics of Program Evaluation," *unpublished manuscript, department of economics,* Harvard University, 2007.

Imbens, Guido, "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review", UC Berkeley, and NBER, May 2003.

Imbens, G. W. "The role of the propensity score in estimating dose-response functions. *Biometrika 83, 706-710, 2000*.

Imbens, G. "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, Papers and Proceedings, 2003, 93, pp.126-132.

Kapoor, A. G., "Review of impact evaluation methodologies used by the OED over past 25 years", Washington DC, The World Bank, 2002.

Keyzer, M. et al. "The CAP Reform Proposal of the Mid-term review", *Centre for World Food Studies*, Amsterdam, 2003.

Kluve et. al, "Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany", *Journal for Labour Market Research*, 2007, 40, pp.45-64.

Kluve et. al, "Assessing the performance of matching algorithms when selection into treatment is strong", *Journal of Applied Econometrics*, 2007, 22, pp. 533-557.

Lechner, M. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review of Economics and Statistics*, 2002a, 84, pp. 205–220.

Lechner, M. "Some Practical Issues in the Evaluation of Heterogenuous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society, Series A*, 2002b, 165, pp. 659–682.

Leuven, E. and Sianesi, B., (2009), PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Statistical Software Components, Boston College Department of Economics, http://econpapers.repec.org/RePEc:boc:bocode:s432001, 2009.

Lokshin, M. and Yemtsov, R . "Has Rural Infrastructure Rehabilitation In Georgia Helped the Poor?" *The World Bank Economic Review*, 2005, 19(2), pp. 311-333.

Mantel, N. and Haenszel, W. "Statistical aspects of the analysis of data from retrospective studies of disease". *Journal of the National Cancer Institute*, 1959, *22*, pp. 719-748.

Michalek J. and N. Zarnekow. "Construction and application of the Rural Development Index to analysis of rural regions in Poland and Slovakia",ADVANCED-EVAL Working paper, University of Kiel, pp. 1-89, revised-version August, 2009.

Michalek J. and N. Zarnekow. "Application of Rural Development Index to Analysis of Rural Regions in Poland and Slovakia", in: Social Indicators Research, 2011, January 29, pp. 1-37, http://www.springerlink.com/content/7844pk504v242552/fulltext.pdf

Michalek J. and N. Zarnekow. "Construction and application of the Rural Development Index to analysis of rural regions". JRC Scientific and Policy Reports 2012, European Commission, pp 1-90.

Michalek, J. (2008): Development and application of advanced quantitative methods to ex-post evaluation of EU RD and structural programme. Presentation at the IV Evaluation Conference on 17.10.2008 in Warsaw/Poland.

Michalek, J. (2008): Measurement of the quality of life in rural areas. Derivation and application of the Rural Development Index to evaluation of EU RD and structural programmes. Presentation at the Polish Academy of Sciences on 20.10.2008 in Warsaw/Poland

Michalek, J. "Application of the Rural Development Index to evaluation of the impact of SAPARD programme in Poland and Slovakia", *paper at international workshop on "Evaluation and Modelling of Rural Development Policies: Theory and Application"*, Kiel 13-14 July, 2009.

Newman, J. et al., "An Impact Evaluation of Education, health, and water supply investments by the Bolivian Social Investment Funds", in: *The World Bank Economic Review*, 2002, 16(2), pp. 241-274.

Ravallion, M.,"Evaluating anti-poverty programs", World Bank, Policy Research Working Paper Series 3625, 2005.

Rawlings, L. B. and Schady, N. R. "Impact evaluation of social funds: an introduction," *The World Bank economic review*, 2002, 16(2), pp. 213-217.

Rickman, D. S. and Schwer, R. K. "A Comparison of the Multipliers of IMPLAN, REMI, and RIMS II: Benchmarking Ready-Made Models for Comparison." *The Annals of Regional Science, 1995,* 29(4), pp. 363-374.

Rosenbaum, P. R. and Rubin, Donald B. "The central role of the propensity score in observational studies for causal effects." *Biometrika,* 1983, 70(1), pp. 41-55

Rosenbaum, P.R. and Rubin, D. B. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome", *Journal of the Royal Statistical Society*, 1983b, 45, pp. 212.

Rosenbaum, P. R. and Rubin, D. B. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score", *The American Statistician*, 1985, 39, pp. 33 - 38.

Rosenbaum, P. R. "Observational Studies", New York: Springer-Verlag. 2nd Edition, 2002.

Rosenbaum, P. R. "Design Sensitivity in Observational Studies", *Biometrika*, 2004, 91, pp. 153-164.

Roy, A. D. "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 1951, New Series, 3(2), pp. 135-146**Rubin, D. B.** "Characterizing the Estimation of Parameters in Incomplete Data Problems", *The Journal of the American Statistical Association*, 1974, 69(346), pp. 467- 474.

Rubin, D. B. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 1974, 66(5), pp. 688 – 701.

Rubin, D. B. "Bayesian inference for causality: The importance of randomization." *Proc.Social Statistics Section,Am. Statist. Assoc.*, 1975, pp. 233-239.

Sianesi, B. "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s," *The Review of Economics and Statistics*, MIT Press, 2004, 86(1), pp.133-155, 09.

Smith, Jeffrey A. " A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies" *Schweiz. Zeitschrift für Volkswirtschaft und Statistik*, 2002, 136, pp. 1-22.

Smith, J.A. and Todd, P. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," CIBC Human Capital and Productivity Project Working Papers 20035, University of Western Ontario, CIBC Human Capital and Productivity Project, 2003.

Smith, Jeffrey A. and Todd, Petra E. "Does matching overcome LaLonde's critique of nonexperimental estimators? " *Journal of Econometrics*, 2005, 125(1-2), pp. 305-353.

Todd, P., "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated, Handbook of Agricultural Economics", Volume 4, North Holland, (eds.) R. E. Evenson and T. P. Schultz, 2006.

Toulemonde, J. et. al. "Three layers of quality assurance: would this help provide EU policy makers with the evaluative information they need?", *paper to the EES Conference*, Seville. (www.europeanevaluation.org), 2002

Treyz, F. and Treyz, G. " Evaluating the Regional Economic Effects of Structural Funds Programs Using the REMI Policy Insight Model", *Regional Economic Models, Inc. Challenges for evaluation in an Enlarged Europe,* Budapest, June 26-27, 2003, ec.europa.eu/regional_policy/sources/docconf/budapeval/work/treyz.doc, 2003.

Van de Walle, D., and Cratty, D. "Impact Evaluation of a Rural Road Rehabilitation Project." World Bank, Washington D.C. Processed, 2002.

Vanhove, N. "Regional Policy: A {European} Approach", Aldershot, UK, 1999.

Van de Walle, D. and Cratty, D. "Is the emerging non-farm market economy the route out of poverty in Vietnam?," *The Economics of Transition*, *The European Bank for Reconstruction and Development,* 2004, 12(2), pp. 237-274.

Winship C.and Morgan C. "The estimation of causal effects from observational data." *Annual Review of Sociology*, 1999, 25, pp.659–707.

Zhao, Z., "Using Matching to Estimate Treatment E®ects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," *The Review of Economics and Statistics, 86(1), 91-107, 2004.*

Appendix 1

Poland: Balancing tests for covariates in GPS

Treatment Interval No 1 - [1.00000000363e-15, 4.965654373168945]			
	Mean Difference	Standard Deviation	t-value
f1	-.14127	.11216	-1.2596
f2	.0201	.02563	.78431
f3	-.26497	.13206	-2.0064
f4	-.19151	.12583	-1.5219
f5	-.01984	.13084	-.15161
f6	-.01833	.01651	-1.1103
f7	-.20739	.14312	-1.4491
f8	.20638	.11484	1.797
f9	-.13374	.12726	-1.0509
f10	-.21071	.12472	-1.6895
f11	-.06286	.12064	-.52105
f12	-.01814	.12029	-.15079
f13	.1841	.11505	1.6002
f14	.15537	.08764	1.7728
f15	.0269	.12543	.21443
f16	.12798	.12965	.98708
f17	.20627	.13674	1.5085

Treatment Interval No 2 - [5.015648365020752, 9.84581184387207]			
	Mean Difference	Standard Diviation	t-value
f1	-.0318	.12339	-.25775
f2	-.00334	.02594	-.12884
f3	.18051	.13998	1.2895
f4	.02207	.13185	.16742
f5	.03903	.1359	.28719
f6	.00401	.0177	.22643
f7	.16831	.14548	1.1569
f8	.05505	.12947	.42523
f9	.05248	.13558	.38705
f10	.12919	.13066	.98876
f11	.01446	.12162	.11886
f12	.12691	.12193	1.0408
f13	-.12492	.13038	-.95814
f14	-.11106	.09373	-1.1849
f15	-.01288	.12839	-.1003
f16	.03126	.1318	.2372
f17	-.01337	.14307	-.09343

Treatment Interval No 3 - [10.02807235717773, 19.44757080078125]			
	Mean Difference	Standard Diviation	t-value
f1	.4374	.17706	2.4704
f2	-.01632	.03967	-.41136
f3	.1237	.22515	.54941
f4	.1708	.21091	.80983
f5	-.13138	.17862	-.73554
f6	.00015	.02697	.00572
f7	.14271	.23742	.60109
f8	-.10332	.15665	-.65955
f9	.12969	.22052	.58808
f10	.11252	.20865	.53929
f11	.16123	.19614	.82201
f12	-.29031	.18782	-1.5457
f13	.01869	.19508	.09579
f14	-.00462	.14782	-.03129
f15	0.465	.19969	.23286
f16	-.56217	.19981	-2.8135
f17	-.14514	.21821	-.66515

Treatment Interval No 4 - [20.01595306396484, 43.77790069580078]			
	Mean Difference	Standard Diviation	t-value
f1	.68467	.5414	1.2646
f2	-.15133	.11012	-1.3742
f3	-.00316	.6111	-.00517
f4	.12934	.56834	.22757
f5	1.5849	.54323	2.9176
f6	-.08145	.07371	-1.105
f7	-.35461	.63296	-.56024
f8	.07122	.42482	.16764
f9	.16631	.57918	.28715
f10	.08475	.57915	.14634
f11	.28795	.55982	.51436
f12	-.44719	.5338	-.83775
f13	-.24201	.55149	-.43883
f14	-.04291	.39758	-.10792
f15	.40546	.5592	.72507
f16	.50293	-58308	.86255
f17	-.28478	.61443	-.46349

European Commission
EUR 25419 – Joint Research Centre – Institute for Prospective Technological Studies

Title: Counterfactual impact evaluation of EU rural development programmes - Propensity Score Matching methodology applied to selected EU Member States.
Volume 2: A regional approach

Author: Jerzy Michalek

Luxembourg: Publications Office of the European Union

2012 – 79 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424

ISBN 978-92-79-25678-3

doi:10.2791/8228

Abstract

The basic objective of this study is to analyze the impact of EU RD programmes on rural regions. Aggregated effects of a given RD programme at regional levels are estimated using the Rural Development Index (RDI) – a proxy describing the overall quality of life in individual rural areas. The weights of economic, social and environmental domains entering the RDI index (composite indicator) are derived empirically from the econometrically estimated intra- and inter-regional migration function after selecting the "best" model from various alternative model specifications (e.g. panel estimate logistic regression nested error structure model, spatial effect models, etc). The impacts of individual RD measures are analysed by means of a counterfactual analysis by applying combination of the Propensity Score Matching (PSM) (e.g. Kernel matching) and difference-in-differences (DID) methods (i.e. by comparing supported regions and matched control group, prior to the programme and after it). Evaluation of programme effects (by programme measures) at regional level is carried out on the basis of the estimated policy parameters: Average Treatment Effects (ATE), Average Treatment on Treated (ATT) and Average Treatment on Untreated (ATU) effects by using the RDI Index and unemployment ratios as impact indicators. Given information on regional intensity to programme exposure (financial input flows by regions) the overall impact of obtained support via a given RD programme is estimated by means of a dose-response function and derivative dose-response function within the framework of a generalized propensity score matching (GPS). Furthermore, sensitivity analysis (Rosenbaum bounds) is carried out in order to assess a possible influence of unobservables on obtained results (under a binary PSM methodology). Above methodologies are empirically applied to evaluation of the impact of the SAPARD programme in Poland and Slovakia in years 2002-2005 at NUTS-4 level. Results show a full applicability of proposed approach to the measurement of the impact of rural development and structural programmes.

LF-NA-25419-EN-N

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.

Publications Office

ISBN 978-92-79-25678-3