WIRTSCHAFTSUNIVERSITÄT WIEN

DIPLOMARBEIT

Titel der Diplomarbeit:

The Regional Dimension of European Framework Programme Research Collaboration: A Gravity Model Approach

Verfasserin/Verfasser: Zeugner, Stefan Klaus

Matrikel-Nr.: 9851051

Studienrichtung: J 140

Beurteilerin/Beurteiler: Hackl, Peter

Ich versichere:

dass ich die Diplomarbeit selbstständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe.

dass ich dieses Diplomarbeitsthema bisher weder im In- noch im Ausland (einer Beurteilerin/ einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe.

dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Datum   Unterschrift

Stefan Zeugner
9851051
stefan.zeugner@cemsmail.org

The Regional Dimension of European Framework Programme Research Collaboration

A Gravity Approach

Diploma Thesis with the Department of Mathematics & Statistics
In Cooperation with Austrian Research Centers Seibersdorf
Wirtschaftsuniversität Wien – Vienna University of Economics & B.A.
March 2005

Supervisor: Univ.-Prof. Dr.h.c.Dr. Peter Hackl
Supervisor from the part of ARCS: Dr. Thomas Roediger-Schluga

Abstract:

This diploma thesis investigates research collaboration under the 4th European Framework Programme among European Union regions: The central piece of data is a matrix of agglomerate collaborative links between 68 "NUTS-1" regions during 1994-1998. After a literature review and descriptive analysis, the matrix is estimated in an exploratory manner, along the guidelines of the gravity model concept. The main findings are that a region's involvement into Framework Programme collaboration depends positively on its number of research personnel, but also on its relative importance on the national level. Ceteris paribus, two region's bilateral collaboration intensity is decreased by dissimilarities in research sector structure, but increased by 10% if both regions belong to the Romanic-language area. Several regions are fundamentally under-represented in the Framework Programme (in particular Eastern German regions) or collaborate more than is to be expected (e.g. Athens).

KEYWORDS: Framework Programme, research collaboration, RJV, regional, EU, gravity model

JEL CODES: C21, O33, O38, R12

Stefan Zeugner
9851051
stefan.zeugner@cemsmail.org

Die regionale Dimension der Forschungskooperation im Europäischen Rahmenprogramm

Ein Gravitationsansatz

Diplomarbeit an der Abteilung für Wirtschaftsstatistik
In Kooperation mit Austrian Research Centers Seibersdorf
Wirtschaftsuniversität Wien
März 2005

Betreuung: Univ.-Prof. Dr.h.c.Dr. Peter Hackl
Betreuung seitens des ARCS: Dr. Thomas Roediger-Schluga

Abstract:

Diese Diplomarbeit untersucht Forschungszusammenarbeit unter dem Vierten Europäischen Rahmenprogramm zwischen Regionen der Europäischen Union: Die Datenanalyse konzentriert sich auf eine Matrix der kollaborativen Verbindungen zwischen 68 „NUTS-1“-Regionen zwischen 1994 und 1998. Auf einen Literaturüberblick sowie eine deskriptive Analyse folgt eine explorative Schätzung der Matrix anhand eines Gravitionsmodells. Die zentralen Ergebnisse besagen, dass das regionale Engagement in Rahmenprogramm-Kollaboration positiv von seiner Anzahl an Forschern sowie von seiner relativen Bedeutung auf nationaler Ebene abhängt. Einerseits wird die bilaterale Kollaborations-Intensität durch Disparitäten in der Struktur der Forschungssektoren vermindert, steigt andererseits aber um 10%, wenn beide Regionen dem romanischen Sprachraum angehören. Verschiedene Regionen sind im Rahmenprogramm fundamental unterrepräsentiert (besonders ostdeutsche Bundesländer) oder wesentlich stärker involviert als zu erwarten wäre (z.B. Athen).

STCHWORTE: Rahmenprogramm, Forschung, RJV, regional, EU, Kooperation, Forschungszusammenarbeit, Gravitationsmodell

JEL CODES: C21, O33, O38, R12

ACKNOWLEDGEMENT[1]

Nach einiger Verspätung (genauer gesagt: ein Jahr nach Plan) kommt nun meine Diplomarbeit und damit mein VWL-Studium zum Ende. Ich möchte diese Gelegenheit nutzen, um all jenen, die mich in meiner Diplomarbeit und meinem Studium unterstützt haben, meinen herzlichsten Dank auszusprechen. Zuallererst möchte ich meinen Eltern *Klaus und Marianne Zeugner* danken, die nicht nur mich, sondern auch mein Studium erst ermöglicht haben. Darüber hinaus möchte ich auch meinen Großeltern, *Marianne und Max Lemberger*, sowie *Hans und Hertha Rohatschek* für ihre Unterstützung danken – ebenso wie dem Rest meiner großen Familie (insbesondere meiner Schwester *Julia*).
Besonderen Dank gilt auch meiner Traumfrau *Katharina Roth*, die mein Leiden und den Druck aufgrund dieser eineinhalb Jahre dauernden Diplomarbeit mit viel Geduld und Einsatz ertragen hat: Nun ist es endlich vorbei (zumindest Leiden und Druck wegen der Diplomarbeit)! Vor allem aber danke ich Dir, Katharina, für die viele Liebe, die mir während der langwierigen Arbeit an dieser Diplomarbeit von Dir zuteil wurde.

Dass ich zu meiner Berufung als Voixwirt gekommen bin und mich darin derart vertieft habe (damit ist nicht nur die Studiendauer gemeint!), ist vor allem dem ausgezeichneten Studienumfeld zu verdanken, dass ich in Wien genossen habe.[2] Dies ist vor allem meinen Studienkollegen zu verdanken: Meinem langjährigen Zimmergenossen und Freund *Andreas Achleitner* – aber fast noch mehr meinen „Fellow Voixwirts“ *Adi Buxbaum* und *Markus Eller*, sowie dem Rest der `98er Partie: *Florian Schwarzlmüller*, *Sepp Meichenitisch*, *Marc Schimpel*, *Phillipp Schweizer*, *Klaus Vondra*, u.v.a.m. Darüber hinaus danke ich auch noch *Jean-Christophe Curtillet*, der mir während der Dauer meiner Diplomarbeit als Chef in meinem Job die benötigte großzügige Flexibilität ohne weiteres gewährt hat.

Schlussendlich möchte ich in bezug auf die technischen Aspekte dieser Diplomarbeit *Dr. Thomas Roediger-Schluga* danken, der mich von seiten der ARC Seibersdorf mit den nötigen Daten versorgt und konstruktiv zum Literaturteil unterstützt hat; sowie *Univ.-Prof. Dr.h.c.Dr. Peter Hackl* für seine prompte und unkomplizierte Unterstützung..

Stefan Zeugner, Bochum 2005

[1] Since the acknowledgement section is dedicated to persons unanimously speaking German, I preferred to write this section in German. During the rest of this study I, will refrain from using “I”, but instead refer to me as “we” or “us”, since I dislike the somewhat presumptuous notion of “I”…

[2] Dieser Satz bezieht sich in keinem Fall auf „Weltklasse-Unis“ oder ähnliches, sondern auf mein privates Umfeld.

TABLE OF CONTENTS

ACRONYMS AND ABBREVIATIONS

For NUTS-1 identifiers, please refer to Table A.1 on p. 179

For *VARIABLE* identifiers (in italic capital letters), please refer to Table A.6 on p. 184

Abbreviations

b€	billion Euro / ECU
m€	million Euro / ECU
p.	page
resp.	respectively
vs.	versus

General Acronyms used in the text[3]

CIS	Community Innovation Survey
CERN	Centre européen pour la Recherche Nucléaire [European Centre for Nuclear Research]
DG	Directorate-General (Top administrative entity of the European Commission)
EC	European Community / European Communities
EPO	European Patent Office
ESA	European Space Agency
Eureka	Not anymore an acronym. Originally: European Research Co-ordination Agency
FP	European Framework Programme
HRST	Human Resources in Science and Technology
ICT	Information and Communication Technologies
IO	Industrial Organisation
IPR	Intellectual Property Rights

[3] Terms in [] brackets denote translations by the author.

KIS	Knowledge Intensive Services
MITI	Ministry of International Trade and Industry (Japanese)
NACE	Nomenclature générale des activités économiques [General Industrial Classification of Economic Activities]
NUTS	Nomenclature des Unités Statistiques Territoriales [Nomenclature of Statistical Territorial Units]
OLS	Ordinary Least Squares
PPP	Purchasing Power Parity
R&D	Research and Development
RJV	Research Joint Venture
RTD	Research & Technology Development
SCI	Scientific Citation Index
SME	Small and Medium Enterprises
WLS	Weighted Least Squares

Acronyms related to the Framework Programmes

APAS	Actions de Préparation, d'Accompagnement et de Suivi [preparing, accompanying and follow-up measures];
ECSC	European Coal and Steel Research
ESPRIT	European Strategic Programme for Research and Development in Information Technologies
FP	Framework Programme
FP4	Fourth European Framework Programme
JOULE	Joint Opportunities for Unconventional or Long-term Energy supply
SPRINT	Strategic Programme for Innovation and Technology Transfer
THERMIE	Technologies Européennes pour la Maîtrise de l'Energie [European Technologies for the Mastership of Energy]

1 INTRODUCTION

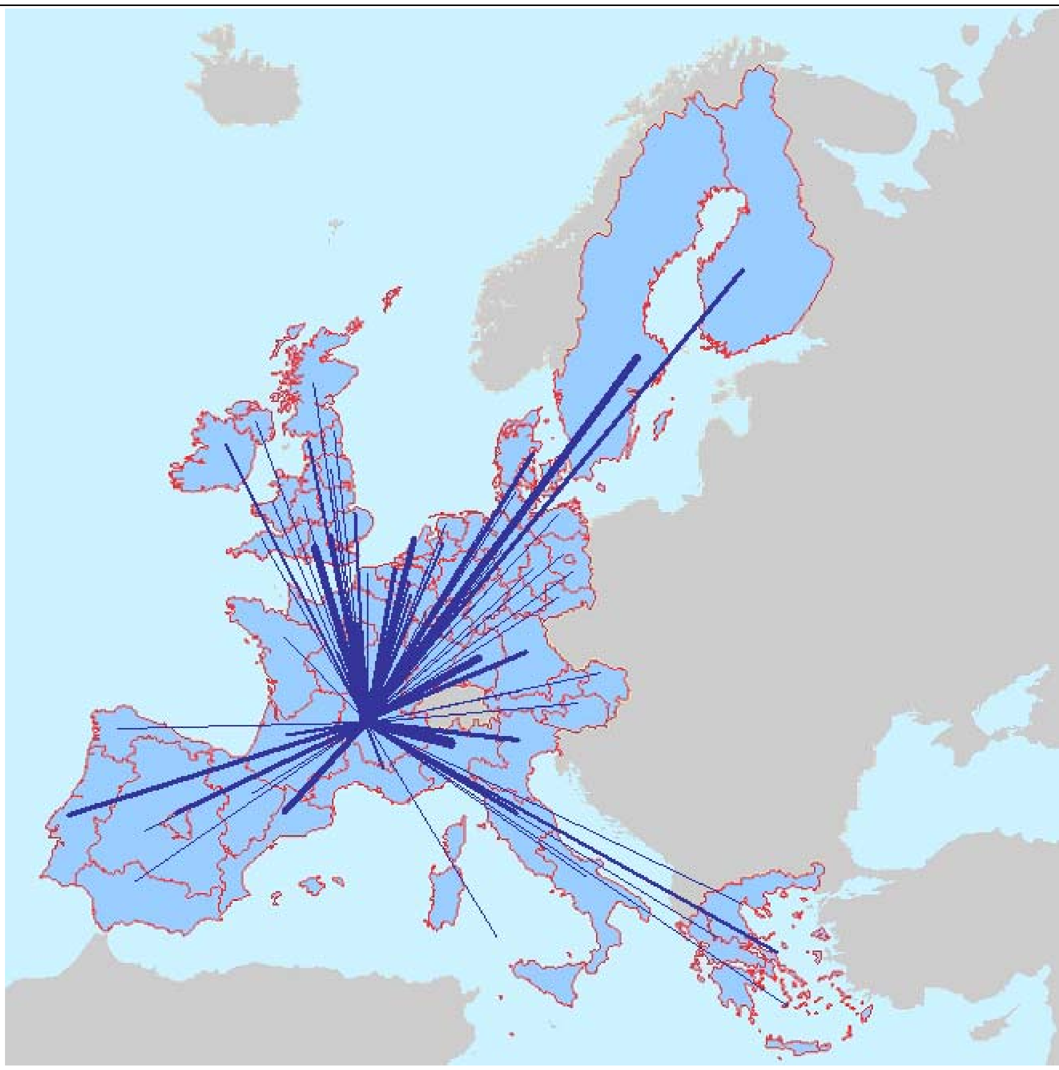
This diploma thesis focuses on the empirical analysis of regional research collaboration under the European Framework Programme. The central piece of data is a matrix depicting research collaboration among 68 EU regions, drawing on newly compiled data by the Austrian Research Centers (ARCS 2003). The primary aim of this exploratory study is to investigate the matrix's characteristics and to identify its potential determinants by relying on the use of a gravity model.

For our purpose, we define the term "research collaboration" rather narrowly as a formal agreement between two or more organisations, directed towards the common achievement of a clearly defined, concrete research objective. Several governmental research collaboration schemes aim at fostering projects meeting these criteria, one of the most prominent among them being the "Framework Programme":

The European Framework Programme (FP) is the broad term for the European Union's research and technology development schemes, administered and funded by the European Commission. Since 1984, EU institutions embed their general research strategy for a predetermined number of years in a "Framework Programme" – this study focuses on the fourth of these consecutive Framework Programmes (FP4), which was carried out between 1994 and 1998. Nearly 80% of FP funds are devoted to promoting research collaboration among public and private organisations in the European Union and beyond. In order to receive public funding for a cooperative research venture, collaborating institutions from at least two EU member states sign up to a common project and file it with the European Commission. Although the FP's budget is less than the expenditure incited by national technology and innovation policies, its international collaborative focus has drawn considerable attention from economic literature: Numerous authors study whether the FP achieves its self-proclaimed objectives (mainly contributing to innovation) and evaluate its scope for technology transfer among European countries. But the major part of empirical research on the subject concentrates on the underlying causes for observed patterns of collaboration. This latter approach is to be followed by the present diploma thesis, albeit from a macro-, agglomerate perspective (see below).

For the study of patterns of cooperative research interaction under the FP, we investigate collaborative links: Empirically, we define such a links to exist between each pair of institutions in a common FP project. Because of this paper's emphasis on the regional dimension, and due to the limited scope of explanatory data, we chose to agglomerate links on the regional levels. I.e. the collaborative links (or “collaborations”) empirically analysed in this thesis are the sums of intra-project links between region pairs. Figure 1 displays a geographic representation of regional collaborative links for the French region Centre-Est (Lyons and surrounding regions).

Figure 1: Collaborative links from Centre-Est to EU-15 regions (FP4)[4]



Consolidating the corresponding data for the 68 regions to be analysed in this study leads to a 68 x 68 symmetric matrix, containing agglomerate collaborative links for each possible pair of regions. This “collaboration matrix” is the focal point and raison d’être for this diploma thesis. Akin to the many micro-level articles to be presented in section 2, we interpret the collaboration matrix as describing a pattern of research interaction. With this respect, the central questions this thesis tries to answer are as follows:

[4] The gauge of the lines radiating out of Centre-Est represents the logarithmic number of collaborative links between Centre-Est and each remaining node.

1. *To which extent do a region's collaborative links within the FP differ from its European counterparts?*
2. *What are the underlying causes for regional differences in FP collaboration?*

The first question refers to the identifying the collaboration pattern captured by FP4 data, as well as putting the data into meaningful representations; which concerns in particular the collaboration matrix and its derivatives (as to be shown in section 4 and 5.1). The second question asks for identifying the factors being relevant to the observed pattern, for their empirical impact on data, and for the economic mechanisms behind the factors' effect on collaboration.

The geographic dimension implicitly determining the regional data, as well as the properties of the collaboration matrix provide the reasons for interpreting the matrix as a pattern of spatial interaction. The collaborating regions are perceived as nodes in a network whose quantity of collaborative links depends on one or several node-specific factors. The intensity of interaction between those nodes is not only influenced by their absolute propensity to collaborate, but also by the "distance" between two nodes which determines the relative inclination to cooperate.

This geographic perception of collaboration data rather resembles the issues investigated in regional economics or economics of trade, as compared to the micro-approaches followed in research collaboration literature. Both regional and trade economics mainly rely on a class of models dubbed "gravity models" (see Haynes/Fotheringham 1988, and Sen/Smith 1995). Basically, this archetype of spatial interaction model is much akin to the Newtonian model of gravitation: The node-specific effect affecting a region's propensity to collaborate may be interpreted as a Newtonian "mass", while the inter-nodal factors affecting relative collaboration intensity[5] correspond to the role of distance in Newton's concept. Mass and distance differ from each other with the first being constrained to one dimension (namely mass) while the second is derived from the positioning of nodes in a space of several dimensions. Regarding the regional character of the data set, we judge the mainstream approach of geography-related economics to be the one suited best to answering the research questions introduced before.

[5] Note: The term „collaboration intensity" will re-occur with various aspects during the thesis. In general, its meaning is the ratio of collaborative links to node-specific properties or factors. As far as "relative collaboration intensity" is concerned, this refers to the number of inter-nodal collaborative links with respect to factors promoting or impeding inter-nodal interaction (such as geographic distance.

The gravity approach offers interesting findings of relevance to both research questions: First data may be classified along, and separated into, “mass” and “distance” effects, facilitating the analysis of the nodes’ general importance, and their relative importance to each other. Moreover, the gravity framework allows for an empirical estimation of the impact by several “mass” (such as research staff”) and “distance” variables (“such as geographic distance”).

The modelling procedure constitutes the central part of this diploma thesis, and is to be carried out in an exploratory manner. The empirical analysis is to be preceded by a review of relevant economic literature on research collaboration in order to identify potential explanatory factors (masses and distances for the gravity model).

Due to the numerous pitfalls in exploratory analysis that may lead to skewed results, we try to apply a balanced view on the prospective determinants’ potential effects. Inter alia, this implies omitting hypotheses on factor impact and importance: Attention is rather shifted to numerous checks for the consistency of results, which should ensure the estimation result’s general validity. The need for cross-checking provided a major incentive for the inclusion of two independent approaches to explanatory factor evaluation: First, the “masses” and “distances” derived from matrix transformation are estimated separately and subsequently re-combined. The second approach attempts to estimate the collaboration matrix directly.

The thesis is thus organised as follows: Part I, the literature review, considers relevant theoretic and empirical literature on research collaboration in general (section 2), and on the institutional characteristics of the Framework Programme in particular (section 3).

Part II is dedicated to analysing FP4 collaboration data by the means of the gravity concept: Section 4 starts with the basic characteristics of the central data set and summarises the basic assumptions and implications of the gravity model. Subsequently, section 5.1 is reserved for a descriptive analysis of collaboration data, while sections 5.2 and 5.3 introduce the prospective explanatory factors for mass and distance, respectively. Section 6 presents the exploratory estimation technique used, and continues with modelling along the two approaches outlined before.

Finally, results are to be consolidated in the concluding section 7. In particular, it checks whether the two different methods applied yield similar outcomes, and whether found determinants exhibit similar behaviour. The discussion of factors’ relative importance is complemented by an intuitive interpretation of empirical results.

PART I – LITERATURE REVIEW

2 THE RATIONALE FOR RESEARCH COLLABORATION

During the last decade, substantial research has been undertaken in order to determine the propulsive and impeding factors for research collaboration, as well as the choice of project partners. This study concentrates on research collaboration within the European Framework Programme (FP). In brief, the FP can be characterised as an international or supra-regional research collaboration scheme. Therefore we omit the vast body of literature focusing on intra-regional collaboration and innovation.

The thus narrowed literature concentrates on supra-regional collaboration in the research sector, originates from several strands of the economics discipline and varies considerable in focus and methodical techniques. The major part concentrates on explaining collaboration between specified types of institutions from a micro-perspective. The corresponding econometric results aim at modelling networks whose nodes consist of individual organisations.

The nodes decide individually whether to collaborate with others and with whom to collaborate. The main part of literature's economic reasoning is directed at understanding these two decisions. This form of reasoning already divides the mechanisms at work into two types: First, a set of factors stimulating or impeding an individual institution's *propensity to collaborate*. Second, factors determining *partner choice*.

In contrast to the majority of research collaboration papers, our approach is based on a regionally agglomerated data set. Its nodes are regional agglomerates – and their interactions are the sum of cross-border relations between individuals. In analogy to literature, we cluster the mechanisms affecting this network into two kinds of factors: First, we assume node-specific factors shaping a region's endogenous potential for collaboration; second, additional factors acting as impediments or promoters of inter-nodal collaboration between regions. We integrate those two classes into a gravitational model; hence we call the former category "mass factors" and the latter "distance factors".

The mass-distance division is one dimension structuring our literature review. Our macro approach demands the introduction of a second dimension – "*micro*" versus "*environmental*" factors. The "micro" category comprises the variables specific to individual organisations: number of staff, or research orientation for instance. "Environmental factors" denote aspects not scalable within the organisation's sphere, e.g. geographic distance or cultural factors.

In literature the distinction between the independent collaboration decision on the one side and partner choice on the other is mostly developed out of a single micro context. Therefore, the section on "micro" factors will be organised along those thematic complexes: Both decisions are found to be related to firm size, explanations for we structure into the "Absorptive Capacity", the "Research Intensity" and the "Basic versus Applied Research" nexus. Furthermore, funding prospects and informal personal contacts are of importance.

The latter theme lays the base for the section on "environmental" factors, which will basically endeavour the patterns shaping the potential for personal interaction, respectively partner choice. Node-specific, independent environmental factors will conclude this section.

2.1 Theoretic and Empirical Reasons for Collaboration in the Microeconomic Sphere

The main part of literature on research collaboration can be categorised along three focal points: Collaboration between firms ("*intra-private*"), collaboration between public institutions, mainly universities ("*intra-public*") and *public-private* cooperation.[6] Papers on intra-private cooperation concentrate on formal research joint ventures (RJVs) between firms and comprise the most influential and the earliest contributions.[7] This strand also disposes of highly elaborated analytical models of intra-private collaboration, mostly originating from Industrial Organisation economics. Extensive statistical enquiries have followed. In the light of the focus on the complex issue of innovation, several authors opted for exploring reality by the means of surveys and case studies.

In the late 1990s, a large number of studies dedicated to the Framework Programmes often led to combinations of intra-public or public-private argumentations, although they mostly concentrated on either of the two viewpoints. Moreover, research on the FP led to the development of large-scale databases focusing on intra-private collaboration: Caloghirou/Constantelou/Vonortas (2001, pp. xxvi-xxxi) provide a comprehensive overview

---

[6] Discussion of private non-profit groups is omitted here for their minor importance in the FP. Moreover the distinction between private and public in this case is not based on ownership but on overall goals of organisations: We generalise those institutions as "private" who are primarily oriented towards profit generation.

[7] In the 1980s, research collaboration shifted into the perspective of economists with the introduction of formal RJV schemes for US firms. Thus the first papers concentrated on intra-private, intra-US collaboration and shaped the later discussion. Compare Link (1981), Katz (1984), d'Aspremont/Jacquemin (1988), Vonortas (1989) and Link/Bauer (1989).

of important databases and surveys: two of the most frequently cited are the “STEP to RJV” database combines data on firms participating in FPs and the trans-European Eureka programme and its successor “RJV-EPO” enhancing the database with European Patent Office (EPO) data. The National Technical University of Athens and MERIT at the University of Maastricht developed and maintain these databases.

Although these databases are large, their range of firm-specific data is often limited to a few financial and statistical indicators. Authors interested in the impact of “soft” factors (e.g. the success of an RJV or personal contacts prior to its establishment) resort to surveys and interviews of different sample scales. The most comprehensive of these surveys in Europe is the Community Innovation Survey (CIS 2004) Three times since 1991, CIS gathered information from more than 40.000 European firms. Like the FP, it is managed by the European Commission’s R&D-related Directorate-Generals, but concentrates on general innovation-related behaviour of firms rather than on R&D cooperation in particular. Academic literature features several similar surveys on a smaller scale.

Interest in public-private research collaboration was in particular incited by the Framework Programmes, and was explored with vigour from the mid-1990s on. Studies of intra-public collaboration, in contrast, developed out of enquiries into academic communication. The latter two topics resulted into less formalised models; conversely the reliance on statistical analysis and cases studies is stronger. Data on public research organisations’ European research collaboration is less institutionalised then on intra-private cooperation, hence corresponding research relies on smaller data samples and surveys.

Hereafter we will try to condense essential concepts applying to either of both organisational types. (However, we will point to organisational differences where necessary.) The section will partly resort to plain theoretic reasoning, but most of it will synthesise arguments assisted by empirical analysis.

First we will review the formalised models of research collaboration and their underlying motivations, mostly spillover internalisation and cost sharing. Then we will present empiric findings on the effects of the modelled mechanisms.

Subsequently, we group the majority of the remaining arguments into two classes related to the concept and the nature of spillovers: the “absorptive capability/research intensity” concept, and the basic vs. applied research decision. Both nexuses stress the important empirical fact that international research collaboration is related to the size of an organisation; therefore we will review relevant arguments in a separate sub-section. The final sub-sections are dedicated to the issue of informal personal contacts, funding and a brief summary of sector-specific factors.

2.1.1 Theoretical Approaches to Research Joint Ventures

During the 1980s, the topic of research joint ventures (RJVs) suddenly began to emerge in European and North American economic theory and policy. Based on the successful model of Japan, governmental subsidies and industrial thrift aimed at improving the respective global comparative technological standing – which in turn led to a rising number of RJVs. Industrial organisation theory early tried to grasp an understanding by analytical descriptions, culminating in sophisticated models as in Kamien/Muller/Zang (1992), Röller/Tombak/Siebert (1997) or Navaretti et al. (2002). Later on, transaction cost economists contributed their part, and the strategic management literature stressed RJVs as an essential component of a strategic cooperation portfolio. Hernán/Marín/Siotis (2003), Hagedoorn/Link/Vonortas (2000), Caloghirou/Ioannides/Vonortas (2003) provide a comprehensive overview of research collaboration models. The latter two pay great attention to the management literature, whereas the former emphasize "hard" economic models.

Caloghirou/Ioannides/Vonortas (2003) broadly cluster theoretic literature on RJVs into "strategic management" papers and "neoclassic" articles. The arguments cited in "strategic management" literature intersect with the "absorptive capacity" approach introduced in the 1990s. As far as relevant, references are provided in the section on "Absorptive Capacity", p. 26. The "neoclassic" articles divide into those based on the transaction cost approach and into papers attributed to the economics of industrial organisation (IO). The main contribution of the transaction cost approach consists of a proper specification of spillover effects, which are briefly outlined below. Industrial economics literature, in contrast, provides analytical models analysing the interplay of spillover effects, cost sharing and the substitutability of products and resources. Due to IO's impact on the conceptualisation of RJVs, its findings will be presented in more depth.

Transaction costs

Transaction cost economics tries to explain economic organisations by defining inter-human transactions as "contracts", whose execution is associated with certain costs. Organisational structures adapt until the most efficient form of contract prevails, i.e. the sum of transaction costs is minimised. Apart from certain patterns of human behaviour, informational attributes specific to transactions may give rise to transaction costs, notably "asset specificity".

"Transaction costs increase steeply when contracts are incomplete, that is, when they do not specify fully the actions of each party in every contingency" (Caloghirou/Ioannides/Vonortas 2003, p. 549)

As an intangible asset, technological knowledge is particularly hard to specify as the underlying of a research contract. This in turn leads to failures in the market allocation of knowledge. Among the commonly identified market failures rank the following:

1) *Pecuniary spillovers* occur because the price of a "piece of knowledge" is not properly set.
2) *Knowledge spillovers* refer to the transfer of embodied knowledge (not necessarily embodied in a "hard" product) between agents without adequate compensation.
3) *Network spillovers* reflect the complementary knowledge, i.e. when the payoffs to one agent increase through the creation of new technology by another agent (without proper compensation).
4) *Opportunism and uncertainty:* The buyer of a technology needs extensive information on its appropriability. When property rights are not perfectly enforced, a potential seller would hesitate to provide the necessary information in advance of contract completion, since the transfer of information would reduce the technology's value.

Transaction cost economics interprets RJVs as an organisational form aiming at economies in transaction costs. The clearly defined joint framework enables partners to partly overcome the incomplete contracts related to technology. Nevertheless, their limited duration allow firms to extract value from joint knowledge on their own. The transaction costs approach has its critics, however: Among the most prominent figure the inobservability of key parameters and the hypothesis that a static environment will let the most efficient contracts prevail. This contrasts to the highly dynamic environment in research.

Industrial Organisation Models

Industrial organisation (IO) theorists rely on the mathematical formulation of the decision to form RJVs. The contributions can be broadly divided into "tournament" and "non-tournament" models (Caloghirou/Ioannides/Vonortas 2003, p. 547). The former, based on game theory, interpret the R&D process as investment in a Schumpeterian game to obtain monopolistic rents. Accordingly, firms in a specific sector aim at gaining "technological leadership", resulting in an "innovation race". The respective literature body was spurred by the emergence of Schumpeterian growth theory in the early 1990s, as well as by simultaneous trends in management science (Prahalad/Hamel 1990). Nevertheless, its impact on RJV literature is still minor to that of "non-tournament" models: This strand of IO research was initiated in the late 1980s and describes R&D as leading to continuous improvements in production and reductions in cost. In that respect Katsoulacos/Ulph (1998) define research trajectories individual to each firm, which form complementarities in the R&D process. Nonetheless, research trajectories are sufficiently similar to allow for the incorporation of knowledge spillovers. These spillovers were identified as one major reason for research collaboration: Since information is a non-rival good and difficult to price, R&D results may "spill over" to external appropriators without adequate compensation – the R&D investor thus cannot enjoy the full benefits arising from newly created knowledge. A collaboration project serves as a mean to overcome those externalities. With cost/risk reduction, a second major motivation to form RJVs was identified by all IO publications. First, pooling of research efforts may lower the risks related to the uncertain research outcomes. Second, it avoids a duplication of research[8] and thus reduces the necessary effort to accomplish a given research target. Inter alia, Röller/Tombak/Siebert (1997) point out that the spillover effect may raise the joint R&D expenditure in the case an RJV is formed, whereas the pooling effect may lead to savings in resourced devoted to R&D.

In this respect, Röller/Tombak/Siebert (1997) and Navaretti et al. (2002) provide two of the most-compelling *non-tournament models*, both supported by empirical evidence. Therefore we will cite some key findings of their papers as a representation of IO literature on research cooperation.

Both base their analytical framework on an influential model by Kamien/Muller/Zang (1992) but infer partly conflicting conclusions. Drawing on d'Aspremont/Jacquemin (1988), Kamien/Muller/Zang (1992) analyse the case of cooperation and competition in R&D and the product markets separately. For that purpose, they lay out a two-stage game: In the first stage firms decide their R&D investment. In the second stage, firms face competition represented by a Cournot/Bertrand game. Solving the model backwards allows a concentrated formulation of parameter dependence. Apart from standard neo-classical assumptions, R&D investment is assumed to affect marginal costs directly. The payoff, i.e. the profits from stage II less the expenditure in stage I, is affected by spillover externalities: Some degree of the reduction in unit costs achieved to R&D will spill over to competitors and reduce their costs. The increased competition may inhibit R&D investment by an individual firm. However, the increased profits of spillover recipients may be internalised through cartelisation or an RJV and thus increase the overall payoff in spite of intensified competition. The authors conclude that, for a sufficiently high spillover rate, cartelisation/RJV is both individually and socially desirable. Moreover, Kamien/Muller/Zang (1992, p. 1297) adjust for the degree of substitutability between the firms' final products, but do not investigate the impact of product complementarities.

Röller/Tombak/Siebert (1997) investigate that particular effect and augment the model for asymmetries between firms (especially asymmetries in size), as opposed to uniform

---

[8] The avoidance of duplication infers from the spillover effect.

production functions in Kamien/Muller/Zang (1992). Instead of analysing competition and cooperation separately, they lay out a three-stage game: First, firms decide on participating in an RJV. Second, they set their R&D investment and, third, they enter Cournot competition. The Cournot solution explicitly accounts for cross elasticity in the product market. Algebraic transformation points out that complementarities (substitutability) in the product market are related to strategic complementarities (substitutability) in R&D investment. Furthermore, the authors infer that initial asymmetries in marginal costs across firms affect the gains from an RJV, i.e. only firms with higher marginal costs will benefit from cooperation.

According to Kamien/Muller/Zang (1992, pp. 1295-1296), RJVs tend to be formed when spillovers create free-rider effects and when duplicative R&D may be avoided through cost sharing. In addition, Röller/Tombak/Siebert (1997, pp. 5-11) conclude that firms may join forces if they sell complementary products and exhibit symmetrical marginal costs (i.e. size). Subsequently, the model is tested on a corporate data-augmented RJV sample divided into seven broad industrial sectors. All firm pairs are tested whether they participate in a common RJV or not.[9] The outcome strongly supports the hypothesis that differences in total assets (measure of firm size) deter firms from joining an RJV.[10] The potential spillover and cost-sharing effects are analysed via the RJV's impact on R&D investment. Although neither effect is dominant across all industries, the firms in most sectors show reductions of R&D investment in case of cooperation, i.e. cost-sharing is prevalent. An additional analysis shows that cost sharing dominates in RJVs with few members (<8). In contrast, RJVs with many participants mostly exhibit increased firm-level R&D, which points to free-rider effects. With respect to cross-sector cooperation, intra-industry cooperation seems to occur more often than inter-industry RJV formation.[11] Nevertheless, the authors identify collaboration patterns to be positively influenced by product complementarities.[12] As a common feature, the cited complementary pairs often enter vertical relationships, e.g. an electronics producer providing communications devices. The results of Röller/Tombak/Siebert (1997) thus support

---

[9] For estimation purposes, a control group of similar firms not participating in an RJV was created. The exogenous factors include balance sheet and R&D data, as well as sophisticated dummy variables. Two estimations are needed two satisfy statistical properties: R&D investment shifts in case of RJV formation provide an auxiliary equation. This is used to test the binary variable whether a firm participates in an RJV or not.

[10] Based on a different data set, Hernán/Marín/Siotis (2000, p. 87) support this view.

[11] „Industrial Machinery and Equipment“ firms are the most likely to form an RJV in general.

[12] It has to be noted that the broad scope of the meta-sectors used in the study reduces the scope for identifying collaboration between firms with complementary products.

the finding of research cooperation along the supply chain, expressed in many survey data analyses.

Navaretti et al. (2002) propose some thoughts sharply contrasting to Röller/Tombak/Siebert (1997): The former elaborate their framework out of the same context as the latter. But they extend the stages of the game to include a firm-specific setting of a research trajectory, i.e. either the two firms have additive or duplicative[13] R&D paths. Similar to Röller/Tombak/Siebert (1997), the papers allows for complementary or substitutable products. By algebraic deduction from some more constraining assumptions Navaretti et al. (2002) formulate an opposing opinion: Ceteris paribus, firms selling substitute products but disposing of additive R&D resources are more inclined to cooperate than firms with complementary products. The reasoning behind is as follows: If firms produce complementary products, it is in their interest to disclose relevant information to the inter-dependent partner voluntarily. The gain from establishing a formal research partnership aiming at internalising spillovers is low – and Navaretti et al. (2002) assume that firms prefer to conduct their complementary research on their own (particularly is there is no cost sharing potential). In the case of substitutable products, companies tend more towards research partnerships, since the internalisation of unintended spillovers needs contracting. The authors test the probability of forming an RJV couple on the affiliation to 4-digit NACE classifications. Unsurprisingly, the results show that this probability is larger when a prospective pair belongs to the same NACE-4 sector.

The model is extended to allow for the possibility of initial cost asymmetries prior to RJV formation. Algebraic transformation leads to the following corollary on RJV formation: If firms produce substitute products, the impact of an increase in cost asymmetry depends on the research trajectory. If the respective pair produces complementary products, a rise in initial cost asymmetry will promote RJV formation. In the subsequent empirical testing, several symmetry indicators are used, of which only the ratio of total sales is significant. The authors (similar to Röller/Tombak/Siebert 1997) relate this result to the degree of cost symmetry, which they find to follow a hump-shaped form: Up to a certain degree of symmetry (which is more than the average of the sample), symmetry increases RJV participation probability, while high symmetry lowers this probability. In contrast to Navaretti et al. (2002, p. 34), we draw the conclusion that symmetry between firms increases cooperation probability, albeit only to a certain limit.

[13] Additive research paths mean that R&D resources may be combined to achieve synergy. Duplicative paths call for cost sharing (Navaretti et al. 2002).

Tournament models evolved out of another strand of IO models and concentrate on “patent races” with winners and losers. These are organised as one-shot or sequential games of firms usually competing in a Cournot market. Participants in the game spend on research for strategic reasons, i.e. to apply a technology before a competitor does. The winners (often a single firm) earn monopolistic rents for their innovation, or may well license the technology to a loser firm. Several authors (e.g. Martin 1994) conclude that although R&D cooperation may be socially desirable, the winner has no incentive to share information with the loser. Public subsidies may provide the necessary push to collaboration. Katsoulacos/Ulph (1998) investigate an augmented model in which they find that firms cooperate voluntarily if they undertake complementary R&D – whereas firms operating substitutable research may join forces when subsidised. Although the existing literature on tournament models is by far less voluminous than the non-tournament strand, it provides a better framework for the incorporation of uncertainty. The implications for RJV formation, however, differ over the respective publications. Moreover, the complicated form of this model type often does not allow for closed-form solutions and is dependent on simulations. In order to reduce complexity, most papers confine themselves to the analysis of duopolies.

To a lesser extent, these caveats also apply to the non-tournament models. Crucial parameters are seldom observable, and there is no clear-cut effect of variations in them. Instead, results depend heavily on parameter constellations. Hernán/Marín/Siotis (2003, p. 76) mention additional disadvantages: First, innovation in IO is formulated either in discrete leaps or continuous reductions in cost. Therefore, they compare non-tournament models to process innovation and identify tournament models with product innovation. A more reality-oriented combination of the formulation is seldom given and does not provide practical insights. Second, there is no duplication of research efforts before or after an RJV is formed (apart from Katsoulacos/Ulph 1998). Third, competition enters the game either via a duopoly, or, in case of more firms, with symmetric cost structures. This setting allows only for continuing competition or for cooperation among *all* firms under consideration.

Although the models’ virtue is to simplify reality, the extremely complex research subject, particularly the nature of spillovers, hinders their practicability. Nevertheless, four factors common with IO models are generally deemed to be of importance to (private) research collaboration: Cost-sharing, spillover internalisation, complementarities in resources and output, and degree of symmetry between prospective collaborators.

2.1.2 Cost-sharing, Spillovers and Complementarities

Cost sharing, spillovers, complementarities and symmetry are the main determinants for research collaboration as emerging from analytical models. The following pages will be dedicated to empirical results on the former three motivations, while asymmetries will be investigated in the section on organisational size, p. 36. In contrast to the papers presented so far, the studies mentioned henceforth do not explicitly base their specifications on analytical models.

Cost-Sharing

From a neoclassic point of view, the non-rival properties of information imply that no expenses incur through information sharing. This view, of course, neglects the transaction cost of information transfer, as well as the fact that the information transferred may be only partly appropriable by the recipient. The strongest incentive for any firm not to share information is to gain competitive advantage. Thus the decision whether to enter an RJV depends on the savings realised through joining resources compared to the loss in competitive position. Röller/Tombak/Siebert (1997, p. 18) point out that the cost-sharing motivation dominates RJV consisting of few members, whereas it is of less importance to large RJVs (in terms of participants). Moreover, the costs of a research project may surpass the profits if it is to be borne solely by any prospective RJV participant. Therefore cooperation may be the only method to render a cost-intensive project feasible. Intuitively, this may be of special relevance to small and medium-sized enterprises (SMEs), but anecdotal evidence and Röller/Tombak/Siebert (1997, p. 18) suggest that particularly large telecom firms are interested in cost-saving R&D cooperation. Miotti/Sachwald (2003, p. 1492) discover evidence that cooperation among European firms aims more at sharing costs than their collaboration with US partners.

Although the neoclassic models in particular neglect the issue of uncertainty, the motivation for fixed cost reduction is quite similar to that for risk sharing: The R&D process belongs to the most complex issues in modern economies and its exact outcome is both highly specific (thus they may be not appropriable) and extremely uncertain. Therefore the pooling of many research projects may raise the return/risk ratio (Pyka/Windrum 2001, p. 10). “Tournament” IO models point into that direction.

The strategic/real options approach focuses explicitly on uncertainty, but its few papers on research collaboration do not provide clear-cut results: First, there may be uncertainty regarding the technical appropriability of research outcomes – in this case, the reasoning mentioned above applies. Second, the profitability of implementing a new technology in the market may be subject to uncertainty. In this case the investment in R&D constitutes the purchase of a call option on a certain technology. The costly decision to implement this technology may be postponed to a later date. The value of this strategic option depends on parameters similar to other strategic options: The more uncertain the benefits of an implementation, the more the option’s value increases, i.e. the more will be invested in R&D. Consider for example the 3G (UMTS) mobile phone technology: In its beginnings, return expectations were affected by tremendous uncertainty, but telecom firms had to invest in R&D in order to be ready should the market work out. The implications for research cooperation are less clear: While cost sharing may lower the option's initial price, it may also lead to sharing the product market. This in turn may depress profits, which will lower the option's deltas (chances of profit). Which effect dominates depends on parameter setting.

As with R&D cost reduction, the strategic management literature brings forward the argument of core competences: Since setting-up an RJV seems to be associated with substantial fixed costs, this may hinder small firms to engage in numerous collaborations (Hernán/Marín/Siotis 2003, p. 87). Large firms, in contrast, may use the savings drawn from RJVs for other "core" purposes (Porter 1985).

Similar to cost, the duration of an R&D project exerts substantial influence on the return from R&D investments. During recent years firms tried to compress the product development phase in order to raise the interest from amortisation. The need for acceleration may surpass the available resources; therefore RJVs may offer advantages in analogy to cost-sharing partnerships (Caloghirou/Ioannides/Vonortas 2003, p. 553).

Spillover Internalisation and Standards

As has been pointed out before, the complete contractibility of knowledge is nearly unfeasible – thus lowering the prospects for market-based solutions. Research collaborations constitute a tool to internalise unintended spillovers. Spillovers may give rise to market failure, thus cooperation may be a more efficient alternative to stand-alone research. A successful internalisation of spillovers via RJVs therefore is likely to result in higher R&D investment after the RJV is formed. This conclusion establishes the possibility to measure the importance of spillovers to RJV participants, as has been undertaken by Röller/Tombak/Siebert (1997). Together with Kamien/Muller/Zang (1992) and Navaretti et al. (2002), they point out that the pace of spillovers, i.e. how much and how fast knowledge leaks out, is crucial to the choice of collaboration over competition in R&D. Mansfield (1985) contributed one of the most influencing papers on the topic: He constructed data on the so-called "spillover lag" (the speed of innovation diffusion within an industry) for several sectors. However, his data did not provide much explicative value for inter-sector differences in the speed of imitation of new technologies. Hernán/Marín/Siotis (2003, p. 85) and Kaiser (2002, p. 766) tested the indicator's relevance to RJV formation and found it sufficiently significant. They conclude that fast diffusion of knowledge in a sector promotes RJV formation for spillover internalisation.

However, spillover lag data is asserted to suffer from accuracy problems, since spillovers are difficult to measure (Mansfield 1985, p. 222). The notion of a certain sector-wide spillover pace (as assumed by most analytical models) might simply be too crude – literature in 1990s showed that the absorption of spillovers requires substantial efforts and that this absorptive capacity is individual to each organisation rather than to an industry. Section 2.1.3 on absorptive capacity will inquire the issue in more depth.

With respect to spillovers, Hernán/Marín/Siotis (2003, p. 84) explicitly include the degree of *market concentration* in a sector:[14] A less fragmented industry means that potential partners are easier to identify, and will increase the potential to internalise a large part of knowledge spillovers by creating an RJV. Testing the indicator on RJV participation reveals an only weakly significant positive dependence on market concentration. Estimating the impact on RJV participation by firms who have already joined an RJV before, market concentration exerts no influence. The authors presume a non-linear relationship between the two variables, but renounce further attempts to explore it. This outcome may signify that market concentration only help in identification in case where no learning effects from prior collaborations exist.[15] Or it may stem from an industry classification defined in too agglomerated bands. Or it may simply be of lesser importance: Most papers on research collaboration do not mention this particular effect.

IO models view collaboration as a mean to overcome the situation that unintended spillovers may improve a competitor's market position and thus depress the potential payoff for an R&D investor. However, competitors may be hindered to put the acquired knowledge to use via the enforcement of intellectual property rights (IPR). Patenting is the most common mean to avert unintended put to use and monopolise the products resulting of proprietary knowledge. But IPR enforcement works best for products whose embodied knowledge is clearly visible to a referee (a chemical composition, for instance). Pharmaceutical, chemicals or biotech companies may thus enjoy a relatively high degree of IPR protection through patenting (Caloghirou/Constantelou/Vonortas 2001, p. 14), while other sectors can rely less on its monopolisation effect. In particular with respect to cross sector differences, the IPR issue thus affects the propensity for collaboration, the choice of partners and the research focus. Section 2.1.4 will enquire these effects in more detail.

Quite similar to the monopolisation of rents through IPR, the economics of networks emphasise another externality to R&D investment: Due to perfect economies of scale, the

---

[14] Hernán/Marín/Siotis (2003, p. 83) use the Hirshman-Herfindhäl Index in order to represent market concentration.

[15] See section 2.1.3, p. 29

distribution of *standards* for new technologies leads to the prevalence of a single standard among competing specifications. In order to avoid the ruinous “standards wars” of the 1980s, technology leaders increasingly choose to develop joint standards with their European/global competitors. A scheme like the FP offers not only subsidies to the purpose, but also defines the sharing of property rights and obligations.

With regard to competing standards on the market or in the development process, European firms might cooperate among them to gain competitive advantage: either to win an open struggle of opposed standards (as in the successful case of GSM mobile communication technology) or to improve the position in global standards negotiations. Luukkonen (2002, pp. 449-450) illustrates the case of the 3G/UMTS mobile standard, which was formed largely to the desires of European firms due to their unanimous backing of a common characterisation. In the case of the FP, enhancing the competitive advantage of European standards might induce the European Commission to subsidise such ventures.

Complementary Products and Vertical Cooperation

As has been highlighted by Röller/Tombak/Siebert (1997), complementarities in final products may be a major incentive for cooperation in general and R&D collaboration in particular. Navaretti et al. (2002) object that in case of complementary products firms have no interest to withhold cost-reducing information, regardless of the counterpart’s will to cooperate. Hence, a formal collaboration agreement is not necessary.

Although those papers do not consider vertical relationships between firms, the reasoning may be extensible to cooperation along the supply chain. In that respect, various empirical studies (especially surveys) point to the clients’ and suppliers’ importance to firm-level innovation. The production process of vertically related companies is highly complementary, and smooth supply chain organisation renders information transfer vital. Therefore strong incentives push these firms to strengthen their competitive position together via joint research. Caloghirou/Constantelou/Vonortas (2001) conclude from two large-scale surveys that suppliers and clients are the most important external information sources, apart from competitors. According to them, CIS data shows that firms rated “innovative” are among the most active in vertical research cooperation. In contrast, the econometric CIS data analysis by Miotti/Sachwald (2003) leads to the conclusion that low-tech sector firms are more involved in vertical R&D collaboration. Luukkonen (2002, p. 443) emphasises that large firms do more often cooperate horizontally. And from earlier remarks we know that small firms are less likely to enter formal R&D cooperation. How does that fit together?

First, the CIS questionnaire collects data on R&D collaboration in general, from manufacturing firms of all sizes. In most EU member states industrial SMEs play an important role and enhance their position through local clustering (i.e. local cooperation) and flexible adjustment. Often the most innovative SMEs of this kind are found in sectors such as automotive production, textiles, industrial equipment etc. – sectors classified “low-tech” by OECD definition. However, these innovative SMEs are less inclined to participate in international, formal RJVs – which constitute the research topic of this publication.

Apart from external relations along the supply chain, one may think of complementary effects to be at work among the entities of a single corporate group. Up to a certain degree it is reasonable to assume that transaction costs to cooperation within a single organisational grouping are lower than on average. Miotti/Sachwald (2003, p. 1486, pp. 1490-1495) consider this effect in their description of French CIS data: While in the total sample only one third of the organisations cooperate in R&D, 49% of the entities belonging to a business group are involved in R&D collaboration. Consequently, the data is tested on the probability to collaborate and includes among its explanatory factors a binary variable of the value 1 if an entity belongs to a group. The latter variable is found to be positive and highly significant in most test settings, particularly regarding vertical cooperation. (Presumably vertical cooperation among the subsidiaries of a corporation.) However, it has to be kept in mind that CIS data asks for cooperation of all kinds, not necessarily in the framework of formal collaboration agreements.

Regarding cross-border FP collaboration, we draw the conclusion that vertical or business group affiliation may matter predominantly among very large organisations along a supply chain, i.e. in a supplier-customer relationship. Although a formal collaboration agreement may not be necessary on a bilateral basis, it could be needed when a third party collaborates with both entities. Moreover, vertically related firms may convert the collaboration into a formal agreement in order to apply for (FP) subsidies (i.e. collaboration subsidies as a “windfall gain”). In particular, we reckon this to take place in case of closely tied industries – therefore the prevalence of intra-industry collaboration in the data set of Röller/Tombak/Siebert (1997).

2.1.3 The Absorptive Capability / Research Intensity Nexus

The broad notion of absorptive capability denotes an organisation’s ability to appropriate external information and put it to use. With respect to firms, Caloghirou/Constantelou/Vonortas (2001, p. lxii) divide absorptive capacity or capability between the “internal capability” to perform R&D, “organisational capability” to exploit knowledge stocks, and “interacting capabilities” necessary for knowledge appropriation from external relationships. Henceforth we will concentrate on latter: In contrast to analytical IO models, it emphasises the investment into resources in order to appropriate external knowledge.

This relates to spillover effects: Rather than being solely dependent on technology or sector, the pace and quality of spillover appropriation depends on the configuration of the individual recipient. The effective absorption of available knowledge is individually and socially desirable due to the non-rivalry in information. Hence absorptive capability is of major importance to the firm, but even more to public institutions, whose aim is the scientific advancement as such. Moreover, the efficiency in knowledge appropriation determines the ability to benefit from high-level international collaboration.

Private Absorptive Capabilities and Research Intensity

Caloghirou/Constantelou/Vonortas (2001, p. xxiii) describe *absorptive capability* as a key factor to the success of research partnerships. Since the efficiency of knowledge creation depends on smooth absorption of external know-how, absorptive capability and the effectiveness of intra-firm R&D resources are interlocked. Link/Bauer (1989) have shown that there is a positive relationship between research cooperation conducted by a firm, its market share and the productivity of its in-house R&D. This leads to the conclusion that absorptive capacity (efficiency of in-house R&D) raises the propensity to cooperate in research and/or vice versa (compare Hagedoorn/Link/Vonortas 2000, p. 579, and Caloghirou/Constantelou/Vonortas 2001, p. lxii-lxv).

Miotti/Sachwald (2003, pp. 1483) state that

"…*absorption capabilities depend on specific investment, including in particular the existence of an R&D department and enough qualified personnel.*"

Given that these resources exist, the cost to obtain useful results from a research partnership is fairly low, implying higher returns from access to external information. Thus absorptive capacity, i.e. the propensity to cooperate, is positively related to a permanent R&D facility and its features. Veugelers (1997, p. 312) assesses the impact of cooperation on internal R&D performance – the results imply that internal R&D expenditure will rise with cooperation if the respective company disposes of an own R&D department. Luukkonen (2002, p. 447) provides anecdotic evidence that large companies exploit the benefits of collaboration by operating in-house R&D projects parallel to cooperative research.

Hence, the probability of RJV participation increases with the absolute importance of R&D to an organisation, i.e. *organisation size and R&D intensity*.[16] We might add that the existing R&D staff is likely to be trained in absorbing information relevant to the firm's sector. If this

[16] Caloghirou/Constantelou/Vonortas (2001, p. xxii) draw the same conclusion and identify complementarities between internal R&D resources and cooperation.

presumption is correct, the returns from an RJV (and the probability to join it) rise when the cooperation partner operates in a closely related sector.

Miotti/Sachwald (2003, p. 1490) test the relationship between RJV adherence and absorptive capacity/R&D intensity via three dummy variables – in addition to control variables like firm size, etc. The first variable provides information whether the respective firm belongs to a high-, mid- or low-tech sector. Second, a binary factor depicts whether the firm maintains permanent R&D facilities. Third, a ten-degree "science" indicator describes the extent to which a company uses external sources of information.[17] On top of the firm size effect, the authors find all three indicators significant and supportive of their hypothesis, particularly the "science" indicator. Moreover Miotti/Sachwald (2003, p. 1486) provide descriptive analysis of French manufacturing firms: Only one third of "innovative" firms in the sample maintain research cooperation of one or the other kind. In contrast, more than half of the high-tech firms and two thirds of the large companies are involved into R&D partnerships. Hernán/Marín/Siotis (2003, p. 84) and Kaiser (2002, p. 766) include sector-specific R&D intensity (total R&D expenditure in terms of total sales) in their estimation of RJV participation probability. Their results indicate a strongly significant positive impact of R&D intensity. Consequently, the literature emphasises absorptive capacity and the factors related to it as one of the most important causes for research collaboration. Absorption is closely related to (and draws on the same indicators as) the issue of *organisational learning*.

Prahalad/Hamel (1990, p. 80) regard inter-firm collaboration as a means to learn and acquire skills from the research partner. These skills not only refer to technological knowledge, but also to the know-how of whether and how to draw benefits from a research partnership. Hernán/Marín/Siotis (2003, pp. 84-85) state that prior experience from collaboration renders repeated cooperation easier. As a quantitative indicator, they take the number of past collaborations in the relevant RJV programme. The indicator is tested on firm participation probability in Eureka programmes and FP – and it is found to be positive and highly significant. The authors provide two possible explanations: Either there are considerable learning effects associated with participation; or the initial participation covers once-for-all fixed costs, which lowers the marginal costs for joining subsequent RJVs. (In fact, both explanations are different conceptualisations of the same effect.) Pohoryles (2002) emphasises learning effects as well: The majority of analysed organisations knew at least one partner through previous collaboration. Besides, this implies importance to the motivation for partner choice – a topic further discussed in sub-section 2.2.2. In their discussion of CIS data, Caloghirou/Constantelou/Vonortas (2001, p. xlix) state that

[17] Data is drawn from more than 4000 French firms in the CIS-2 survey.

"*Previous experience is by far the most effective way of getting in contact with the most important external source of knowledge […]*"

This points to the importance of organisational learning to absorptive capabilities in general.

Public Absorptive Capability and Scientific Excellence

The notion of absorptive capability holds as well for non-profit organisations. Apart from dependence on the number and excellence of researchers, organisational learning with respect to research collaboration constitutes an important contribution to the enhancement of absorptive capacity (Van der Meulen 2002, p. 342). Analogously to inter-firm collaboration, previous FP participations may provide a proxy variable for learning effects. Furthermore, organisational learning is inter-connected with the topic of contacts and personal networks – see sub-section 2.1.6.

Apart from an organisation's size, its reputation may be raised even more through the excellence and productivity of its researchers. The measurement of academic research productivity has been an important issue to literature throughout the last decade. In most cases, authors focus on bibliometric measures, such as total article output, number of citations and impact factor. For universities, Geuna (1998) hypothesises an impact of research excellence in addition to the pure effect of a department's size. For that purpose, he considers the number of SCI publications divided by staff as an indicator for research productivity and corrects for size effects. The results exhibit a strongly significant influence on both the number of FP participations as well as the decision whether to participate in general.

For non-profit institutions the notion of absorptive capacity and scientific excellence is even more important, since their leading objective is not specifically the conversion of knowledge into profitable products and services, but the generation of useful knowledge as such. Conversely, rewards to public research are related to its excellence and reputation, therefore the increase of recognition might be the prime target of "selfish" public research organisations. From a more altruist point of view, a researcher's motivation may lie in advancing science as such.

In that respect, the INNOCULT survey (Pohoryles, p. 334) depicts "intellectual reasons" as the strongest motivation for joining a European research project. Hakala/Kutinlahti/Kaukonen (2002, p. 371) add that for public institutions the "strengthening of international collaboration" is the top cited benefit from FP collaboration. Thus academic institutions seem to be inclined to pursue international cooperation for an end in itself. This may be partly related to the creation of personal networks among researchers.[18] Hakala/Kutinlahti/Kaukonen (2002, p. 370) infer from the data that academic researchers perceive international cooperation as a means to remain up to date. In an idealistic manner, we suppose that the possibility of new knowledge as such is enough motivation for researchers to engage in collaboration. Interestingly, the respondents to a company survey by Caloghirou/Tsakanikas/Vonortas (2001, p. 159) testify that keeping up with the latest research developments belongs to the most important motivations for public-private collaboration. These results support the view that the enhancement of scientific skills (i.e. absorptive capability) is a major driver for research cooperation, particularly when involving public organisations. Miotti/Sachwald (2003, p. 1486) analyse French data and conclude that scientific cooperation with public institutions is particularly pronounced for large (and patenting) firms. Cohen et al. (1997) provide some further arguments from strategic management, noting that UI collaboration raises firms' sales, patenting activity and R&D productivity. Both statements may not only stem from the immediate research outcome, but also from enhancing the absorptive capacity of a firm collaborating with universities.

International Collaboration as a Result of Research Orientation

Various studies highlight the relation between an organisation inclination towards international research collaboration and its absorptive capability.

Miotti/Sachwald (2003, p. 1492) extend their studies by analysing discrepancies between intra-European and transatlantic RJVs. They test the hypothesis that firms at the "technological frontier" are inclined to cooperate with other leading companies. They hypothesise the majority of those companies to be located in the USA, hence companies interested in the leading edge of a technology should cooperate more with US partners than the average. This implies that "second tier" cooperation aimed at cost cutting should dominate in intra-European R&D collaboration. Consequently, the authors estimate the impact of US competitive advantage in a specific sector and include the R&D intensity and size of a firm. The results demonstrate that firms are more likely to cooperate with US partners the bigger they are in size, the stronger the US competitive advantage and if the firm belongs to the high-tech sector. Mid-tech and mid-sized firms are much more inclined to collaborate with European partners. In addition, Miotti/Sachwald (2003, p. 1495) show that the share of innovative products in turnover depends positively on firm size, market share and cooperation with American partners. Firms constrained to research partners at the

[18] Compare section 2.1.6 on personal motivations, p. 38.

national level are the least likely to produce "innovative products" whereas companies collaborating with European partners take an intermediate position.

This result might arise from the fact that organisations at the technological edge have to go farther to find peers. The domestic region or nation may simply be to small and offer to few prospective partners for specific knowledge generation. Or the spur to learn from the technological leader may incite firms to seek collaboration with the leading country in a technology. However, facing a common threat of strong external competition, R&D-intensive firms may also team up domestically to meliorate their market position. (Eventually, enhancing the strength of domestic firms was a key motivation for the public decision to subsidise RJVs.) In this respect, Link/Paton/Siegel (2002, p. 1463) consider the trade balance in technologically advanced products as a determinant of research partnerships. They find significant parameters supporting the reasoning given above. If external competition steps up, comparative advantages by foreign firms are amplified, and their European followers will even more be willing to choose cooperation partners outside the continent.

Furthermore, markets or research topics may differ in their integration of foreign knowledge. The standards for knowledge transfer may differ among nations, and that to a varying degree across sectors. Or historical developments may incite a sector to be more oriented towards domestic than to external knowledge input (think of Minitel, the French alternative to internet, for instance). The "internationality" of a scientific field shapes the geographic distribution of collaborative links for both private and public research.

Hakala/Kutinlahti/Kaukonen (2002, p. 358) for instance, note that patterns in international academic cooperation differ according to research discipline. "Hard" fields (e.g. technical subjects or natural sciences) study topics of universal relevance and thus tend to be oriented towards international benchmarks. Moreover they are characterised by highly standardised formal languages and specialised audiences, which eases the communication between researchers and bridges differences in cultural context. "Soft" fields (e.g. humanities, social sciences or law), in contrast, lack a standardised nomenclature and often focus on research topics limited to a national audience. Therefore, interaction and subsequent cooperation is reckoned to be less pronounced in "soft" fields. Sandelin/Sarafoglu (2003, p. 4) support this statement: Their paper provides data on publications registered with ISI (a database oriented towards the English language) compared to total population. While non-English-speaking countries hold the highest-ranking positions in natural sciences paper productivity (with the

US at the bottom of the table),[19] English-speaking countries clearly dominate in papers on social sciences, arts and humanities. We presume that this is due to the fact that researchers in the latter fields are more inclined to publish their results in the predominant language of their country. This, in turn, supports the hypothesis of "hard" fields to be more international. Moreover, the English language by itself constitutes a standardised *lingua franca*. Therefore the lack of English knowledge constitutes one of the most important barriers to academic communication and cooperation (Button et al. 1993).[20]

Geuna (1998) tests the probability of university participation in FP projects. In the course of his estimation, he includes the number of scientific disciplines per university as an exogenous factor. The significant parameters for universities oriented towards technical studies (insignificant for natural science, however) point to the assumed impact of technical orientation on R&D collaboration.

2.1.4 Basic versus Applied Research and Intellectual Property Rights

The bias of a research subject towards basic, generic research versus marketable applied research is widely recognised as a major determinant of the decision whether to collaborate as well as for the choice of project partners. In particular, literature uses this distinction to analyse the patterns of public-private collaboration. Since the FP is intended to support "pre-competitive" collaboration, and since public entities account for the majority of project participants, the motivation for cooperation in basic research is of special importance to this paper.

Luukkonen (2002) studies FP participation by Finnish companies, and separates RJVs according to their market orientation as opposed to technology orientation. *Market orientation* denotes RJV participation for the purpose of developing new products, learning about new markets and creating potential marketing alliances. *Technology orientation* derives from the motivation to learn from partners, enhance the knowledge base, monitor the development in the field ("tech watch"), train R&D personnel or develop contacts to major research partners (Luukkonen 2002, p. 441). Technology-oriented research is therefore comparable to generic, pre-competitive research. A descriptive analysis shows that more than 80% of large firms, but only about 30% of SMEs join RJVs for technology reasons. In contrast, market

[19] Remarkably, the ranking of paper productivity in natural sciences is led by Germanic-language countries (i.e. by countries with a large proportion of the population speaking a Germanic language). As a notable exception, Germany and the US rank farther below.

[20] Compare section 2.2.4 on language, p. 45

orientation is cited by more than 85% of SMEs, but less than half of large companies. Larger firms are more inclined to cooperate with competitors, but also fear that sensitive knowledge might leak through the network. Although the European Commission promotes near-to-market research partnerships, large firms are reluctant to conduct market-oriented research in the FP contexts. The FP setting is perceived to promote long-term, risky research. Moreover, the programme's rules require information to be shared with all participants, which further deters market-oriented R&D from FP projects.

For US data, Link/Paton/Siegel (2002) test this finding on a large data set. They draw on Link/Bauer (1989), who assert that due to intellectual property rights (IPR) issues research partnerships will focus on basic rather than applied research. The former define a variable of sector-wide spending on development as a percentage of industry R&D expenditure. A test on the probability to form an RJV confirms the underlying presumption, albeit only at marginal significance.

It is widely accepted that research topics more oriented towards basic research generally favour the inclusion of universities into RJVs. Regarding the FP, universities participate in more than two thirds of industrial RJVs and are even more frequent partners than vertically related firms (Caloghirou/Tsakanikas/Vonortas 2001, p. 158). Their distinct objectives and organisational culture, however, constitute strong impediments to enhanced cooperation. In this respect, a major concern raised by industrial partners is the fear of intellectual property leaking out via universities.

Baldwin/Link (1998) argue that lower appropriability and higher uncertainty of research promotes collaboration with universities. As a proxy, they use the number of RJV participants: Large research partnerships are formed for spillover internalisation, and when product marketability is still far. In this case IPR issues should implicitly be of less relevance.[21]

Hall/Link/Scott (2001, p. 89) investigate the issue by a survey[22] and find that one third of private enterprises consider IPR problems an "insurmountable barrier" to university-industry (UI) collaboration. Subsequently they analyse factors influencing the probability of "insurmountable barriers" to university participation in an RJV and come to some surprising insights: First, public funding increases the likelihood of barriers. Project duration, in contrast, exerts negative influence on barrier probability. Hall/Link/Scott (2001, p. 94) explain this by

[21] Caloghirou/Tsakanikas/Vonortas (2001, pp. 156) illustrate this by calculating a mean of five participants in pure industrial RJVs, whereas partnerships involving universities consisted of eight participants on average.

[22] Firms participating in one of 38 projects within the US Advanced Technology Program.

the two dimensions of "appropriability" and "uncertainty": The less appropriable research results, the less they transform into direct economic gains and the more they are of public nature. It is argued that this renders IPRs harder to define, since universities are inclined to make "basic" research publicly accessible. The conflict of the "two worlds" is reflected in a higher share of public subsidies in a project's funding, whereas firms provide less investment – hence the relationship found in empirical results. Moreover, the authors (Ibid.) conclude that high uncertainty over research results causes a definition of IPR to be meaningless at the start of a project. IPR issues thus will be postponed and resolved at a later stage. The uncertainty factor is captured by the duration of a project, hence its negative impact on IPR barrier probability.

Second, prior experience with university cooperation translates into a positive, weakly significant impact on IPR barriers. This is interpreted as learning effects rendering firms aware of IPR problems. Third, firms in the chemical industry set them distinctly apart by emphasising IPR fears much stronger than in other sectors. This fits well into the pattern that product patenting is of much more relevance to pharmaceuticals/chemicals firms than to the average firm (Caloghirou/Constantelou/Vonortas 2001, p. 14).

Caloghirou/Tsakanikas/Vonortas (2001, p. 157) include both the number of RJV participants and duration as explaining factors for UI collaboration. Both corresponding parameters are found to be significant and positive, while their cross-effect (participant number times duration) is not considerably different from zero. The authors interpret these findings as a strong indicator for universities' importance in basic or generic research.

Caloghirou/Tsakanikas/Vonortas' (2001) estimation results are supplemented by an interview survey: Company respondents testify that keeping up with the latest research developments belongs to the most important motivations for UI collaboration. Consequently, they cite enhancing their knowledge base as the prime benefit from university-industry RJVs. Time-to-market concerns, in contrast, exert strong negative influence. This again confirms the hypothesis that public-private collaboration concentrates on generic research.

Miotti/Sachwald (2003, p. 1491) fit the probability of industrial-public cooperation on several factors, and confirm that a firm's number of staff has a strongly significant, positive impact. Moreover, they include public funding[23] as an exogenous factor and find its effect evenly large, positive and strongly significant. They interpret public funding as an indicator for basic research and thus support the hypothesis of UI collaboration to focus on far-from-market topics. In addition a "science" index and "permanent R&D facility" as indicators for absorptive capacity figures as well among the significant and positive factors (on top of firm size), which

[23] A dummy variable taking the value of one, when a participating firm benefits from public subsidies..

once again highlights the relationship between absorption capabilities and the scientific community. Although Hall/Link/Scott's (2001) conclusions regarding uncertainty assert this opinion, their view of public funding increasing IPR barriers conflicts with the results of Miotti/Sachwald (2003). The positive relationship between public funding and IPR barriers may evenly arise from the firm's weaker standing in IPR negotiations due to their less important share of investment. Moreover, it has to be noted that the authors base their conclusions on a modest sample size.

Furthermore, Miotti/Sachwald (2003, p. 1486) conclude that scientific cooperation with public institutions is particularly pronounced for large and patenting firms. The first characteristic points to absorptive capacities mention in sub-section 2.1.3. Patenting firms are relatively more frequent in the chemical or biotechnology disciplines (Caloghirou/Constantelou/Vonortas 2001), which exhibit higher propensity for UI cooperation. Reliance on patents also decreases the dependence on secrecy. Therefore IPR fears may be less important to patenting firms.

Luukkonen (2002, p. 442) presents the somewhat surprisingly finding that companies in sectors of low R&D intensity are more inclined to embark on technology-oriented projects, whereas high-tech firms dominate near-to-market RJVs. As an explanation, it is presumed that low-tech firms conduct very short-lasting market-oriented research, which is inappropriate to formal FP projects. Miotti/Sachwald (2003, p. 1491) confirm the impact of patenting and absorptive capabilities. Furthermore, they hypothesise that R&D cooperation with universities depends positively on the importance of R&D within a firm's sector. Consequently they test affiliation with high- and mid-tech sectors on UI partnership formation – but the results indicate that high-tech firms are less involved into UI collaboration than their lower-tech counterparts.

Caloghirou/Tsakanikas/Vonortas (2001, pp. 157-158) consider the distribution of academic participants within industrial RJVs according to key action lines. In all but one research topic, universities participate in more than half of total projects. Their share is particularly high in biotechnology (89%), but also in the inter-related fields of agriculture, fisheries, food and environment. The high share in biotechnology may be attributed to the importance of up-to-date scientific research in this sector. Luukkonen (2002) states that biotech firms are often heavily involved in academic research, and frequently evolve out of university spin-off companies. She confirms the observation that pharmaceuticals companies are particularly engaged in UI cooperation.

The share of universities in production-oriented fields (e.g. telecommunications, manufacturing, ICT, aerospace, electronics) is considerably lower. Caloghirou/Tsakanikas/Vonortas (2001) provide no explanation for this result, apart from the determinants for biotech and the agriculture/environment nexus. Possibly these sectors are more dependent on applied engineering as compared to basic research. At least, the respective data may explain the results of Miotti/Sachwald (2003): Apart from biotechnology (whose importance is still minor to electronics and ICT), the mentioned industries constitute the bulk of firms classified as “high-tech”. Companies in the food or agriculture sector, in contrast, are seldom encompassed by this category, hence the negative relationship between “high-tech” and UI cooperation.

2.1.5 Organisational Size and Symmetry

Most of the papers cited so far conclude that firm size raises the propensity to cooperate in R&D.[24] One reason may be that RJVs are often associated with large fixed costs: e.g. the establishment of specific facilities or research teams, incremental assignment of labour to the project, etc. Large firms may easily spread these fixed costs on revenue, thus for them it is less costly to join one or several RJVs. For that reason, it is also more likely that large firms maintain permanent R&D divisions, whereas SMEs set up teams on the occasion. This once again reduces the incremental cost of research collaboration for large firms – and RJVs may also be joined to employ existing R&D resources during weak demand for internal research.

We presume that the capability to communicate in English and other foreign tongues is better developed among large corporations, since they are more oriented towards an international clientele. Furthermore, absolute size should ease the identification of prospective collaboration partners in a complex environment. Small firms, in contrast, are reported to search relatively more often for expertise at the regional level. This may be due to their personal contacts, or to the fact that the regional level simply does not provide sufficient human and capital resources adequate to a large firms needs. Moreover, public authorities may have a preference for “big” or “small” business. If RJVs were only filed for competition law reasons, notifications by SMEs would be improbable.

In total, most authors argue in favour of positive relationship between firm size and R&D collaboration. It has to be noted that the fact of asymmetries in size between potential project partners also plays a role.[25] According to most oligopolistic models, firm size reflects relative efficiency in homogenous product markets. Thus firm size parallels cost advantages in IO models – Röller/Tombak/Siebert (1997) view the two as equivalents. With regard to unintended spillovers taking place in an RJV, an efficient (large) firm may be reluctant to join

[24] Inter alia, Caloghirou/Constantelou/Vonortas (2001), Luukkonen (2002), Hernán/Marín/Siotis (2003) and Miotti/Sachwald (2003)

[25] See section 2.1.1, p. 20.

forces with a less efficient company. Thus the more symmetric two prospective RJV partners are in size, the more they are inclined to collaborate. Röller/Tombak/Siebert (1997) test an indicator of asset size symmetry and obtain results supportive of their theoretical findings. Luukkonen (2002, p. 448) supports this view by asserting that small R&D-intensive start-ups (biotech SMEs) are reluctant to share their knowledge with major firms in their sector. Hernán/Marín/Siotis (2003, p. 84) confirm asymmetries in size as a deterring factor to RJV formation.

In most papers, size is represented by a firm's number of staff, but financial indicators (sales, assets) are sometimes used with better results. *Market share* may as well be related to firm size and reflect efficiency. A large market share may as well render the firm easily identifiable as a leading firm in its specific sector. (Note that firm size and market share is not perfectly equivalent, since large corporations are often involved into several, very heterogeneous markets.) Moreover, market share reflects relative firm size, thus the competitive position of a company. In their estimation of RJV participation, Miotti/Sachwald (2003) include market share and firm size as exogenous variables. The outcome shows that market share has a positive impact on top of firm size. We interpret these results as underlining the importance of relative firm size.

As with inter-firm cooperation, the absolute and relative size of an academic research facility seems to have a considerable impact on research cooperation. First, larger research organisations (in terms of personnel) simply dispose of more potential resources for collaboration. Second, larger entities may achieve economies of scale in the fixed costs associated with research collaboration, and thus dispose of over-proportionally more collaborative linkages. Third, increasing size renders an organisation more identifiable as a potential partner. The size effect may well refer to homogeneous research groups, but the reasoning may also apply to whole organisations:

As Geuna (1998, p. 679) points out, the funding agency (i.e. the European Commission) only has limited information on particular research groups at hand. As a substitute, the Commission draws on the name/reputation of the corresponding university in order to determine a participant's potential. Furthermore, Geuna asserts that well-known names are associated with positive image externalities on less renowned firms and universities. Consequently, one could assume symmetries to play a role in analogy to inter-firm cooperation: If reputation externalities are important, well-known institutions would choose partners maximising their own benefits from image spillovers, i.e. similarly renowned research partners. Less-known organisations would thus be confined to collaboration within their band of reputation.

Survey results analysed by Pohoryles (2002, p. 335) indicate that research cooperation also has positive effects on reputation: Raising the organisational reputation is one of the major incentives for collaboration. The collaborative project's impact on personal reputation ranks as well among the top motivations at the individual level.

Geuna (1998) provides an estimation of the number of staff's impact on the number of university participations in FP projects. The resulting parameter is both large and highly significant. However, Geuna (1998) finds no relationship between size and the decision whether to participate European cooperation programmes in general. In addition, he reckons that reputation may be influenced by the historic achievements of a university, i.e. its age. The inclusion of dummies depicting the period of the entities' establishment into his estimation, however, does not deliver any meaningful result.

Regarding public-private cooperation, large firms are found to be more inclined towards public-private collaboration (Miotti/Sachwald 2003). They may also dispose of greater absorptive capability in order to convert (generic) scientific results into profits.

2.1.6 Personal Contacts and Motivations

As Caloghirou/Constantelou/Vonortas (2001, pp. xxxx-li) point out, personal contacts between researchers are the main channel for flows and absorption of tacit knowledge. In this respect, research collaboration is perceived as only one possible tool to gain access to external knowledge. Informal contacts or the exchange of personnel are at least of equal importance and pave the way for formal collaboration:

A descriptive survey analysis by Pohoryles (2002, p. 334) shows that 71% of analysed entities knew a least one FP partner organisation through previous collaboration, and 58% even knew at least one person at the FP partner from prior cooperation. Furthermore, more than half of researchers surveyed reported to have previously worked for a partner institution. Pohoryles (2002) interprets this fact as an indicator for high labour mobility among institutions involved in the FP. Informal personal interaction thus is viewed to facilitate the decision whether to cooperate, and definitely affects the choice project partners. For our purpose, we acknowledge the importance of personal contacts on an agglomerate basis as the potential for personal interaction between regions (Compare section 2.2.2).

Relations among individuals are also relevant to the nexus of personal networks and research partnerships. Existing social networks seem to be an important basis for the choice of partners and for formal research cooperation. The study of Button et al. (1993) implies that young researchers first focus on building a domestic network, than try to create European connections and at last extend their network to the international level. This may have changed: Hakala/Kutinlahti/Kaukonen (2002, p. 371) hypothesise that FP collaboration may promote younger researchers' careers. Anecdotic evidence supports the presumption that younger and mid-age researchers are more active in maintaining European networks. In particular, this refers to the social sciences and humanities, where only recently an orientation towards international forums has occurred. Young researchers, often with the experience of stays abroad, may seize the opportunity to expand their valuable networks – as opposed to well-established senior researchers with their domestic networks already operating at full scale. We therefore expect a positive relationship between an organisation's share of young researchers and its European collaborations.

Apart from the general motivations mentioned above, individual researchers' propensity for international cooperation depends on the orientation of organisational culture. Van der Meulen (2002) introduces the term „Europeaness" which includes FP participation and funding among its components. Moreover, his survey assesses the individual's orientation to the university organisation as opposed to the scientific community of the relevant discipline. In contrast to his expectations, Van der Meulen (2002, p. 353) finds that researchers who are formally, but above all socially oriented towards the scientific community are more inclined to European collaboration.

2.1.7 Access to Funds and Legal Advantage

Public funding is a key characteristic of the framework programmes. Pohoryles (2002, p. 336) finds evidence that access to funds ranks among the major determinants of European research cooperation. Miotti/Sachwald (2003, p. 1489) include a dummy variable for public funding in their estimation of CIS data. They find it to be strongly significant and positively influencing R&D collaboration. Veugelers (1997, p. 311) argues that public funding of research increases internal R&D rather than leading to crowding-out. Furthermore, she provides evidence for an indirect positive effect on R&D collaboration. Apart from the cost sharing intentions mentioned above, other kinds of external funds provided to a firm may work into a similar direction.

The need for public funds is frequently justified on the grounds of internalising externalities; they therefore enable increased research. The *legal advantages* offered by formal RJVs similarly enable cooperation otherwise obstructed by law. In particular, the easing of competitive law restrictions allows collaboration projects to internalise most of unwanted spillovers in an industry – i.e. to encompass many of the major companies in a sector. An example may be found in large telecom projects aiming at the creation of new communication standards.

External funding is presumably more important to researchers from public institutions than to corporate R&D departments. In their analysis of FP participation by Finnish universities, Hakala/Kutinlahti/Kaukonen (2002, p. 371) find that “funding for new projects” is considered the second most important benefit from cooperation (after “intellectual reasons”). Pohoryles (2002, p. 334) analyses data of the European INNOCULT survey with respect to motivations for creating research networks: The results imply that funding is most important to researchers from universities, and in the Environment programme (a programme particularly dominated by non-firm participants).

It may be of relevance to distinguish between external funding through cooperating firms and subsidies by public institutions. As Caloghirou/Tsakanikas/Vonortas (2001, p. 153) point out, industrial funding of research is rather minor (5% of total funding on OECD average) but increasing rapidly. For our purpose, however, public funding (i.e. FP subsidies) is more important. In this respect, Hakala/Kutinlahti/Kaukonen (2002) identify an inverse relationship between the availability of funds from national sources and the importance of European subsidies. Van der Meulen (2002) stresses that increased dependence on competitive funding is associated with higher importance of European subsidies.

2.1.8 Business Environment Factors

Link/Paton/Siegel (2002, p. 1463) reckon that a *cyclical economic downturn* reduces the firm-level funds available for R&D. Cooperation for the grounds of cost sharing seems to be one of the consequences. Accordingly, the authors test this negative effect and find weakly significant parameters supportive of their reasoning.

Many authors have highlighted *cross-sector differences* in cooperation propensity. Luukkonen (2002) cites evidence that pharmaceutical companies are reluctant to collaborate with competitors for IPR fears. Cooperation between telecommunication firms, in contrast, seems more oriented towards cost-sharing than in other industries, therefore we would expect intra-industry cooperation to be more pronounced in this particular sector. Luukkonen (2002) adds that specific sectors frequently have institutionalised cooperation agreements in place, which provide a platform for horizontal collaboration (e.g. Finnish pulp and metal industries).

Röller/Tombak/Siebert (1997), Navaretti et al. (2002) and Hernán/Marín/Siotis (2003) dedicate large parts of their work to sector disparities due to different diffusion speeds regarding knowledge spillovers. The latter draw on Mansfield (1985, p. 220), who attributes over-proportional diffusion speed to the pharmaceuticals, electric and instruments industries. The latter two comprise the modern information and communication technology sectors: Firms from these sectors are regarded as particularly inclined to cost-sharing RJVs (Röller/Tombak/Siebert 1997 and Luukkonen 2002) although the motivation from spillovers is regarded as evenly elevated. This may explain the high collaboration intensity of the industry on regional and international scale (Caloghirou/Constantelou/Vonortas 2001, p. lxiv).

Caloghirou/Tsakanikas/Vonortas (2001, pp. 157-158) highlight the over-proportional involvement of patenting firms in university-industry collaboration. This relates predominantly to pharmaceutical, chemical and biotech companies, who carry out extensive basic research with public institutions. Fontana/Geuna/Matt (2003, p. 8-11) confirm the importance of public research organisations to chemicals firms. The electronics, information and communication sector, in contrast, was found to patent less due to lower IPR protection feasibility. Conversely, the latter's dedication to public-private and to basic research is found to be less pronounced (Fontana/Geuna/Matt 2003). Hall/Link/Scott (2001) object that for pharmaceutical and chemical firms IPR fears are the most deterring from public-private collaboration. However, they state also that only through previous collaboration private firms become aware of IPR issues.

2.2 Environmental Determinants Affecting Research Collaboration

Literature (inter alia Pohoryles 2002, Caloghirou/Constantelou/Vonortas 2001) stresses the importance of personal networks to research collaboration. With respect to our study, informal personal interaction shapes both the choice of collaborators and the decision whether to cooperate on R&D. The latter effect may be interpreted as a core-periphery factor: The less external personal relations exist, the more a node/organisation is remote to potential collaborations, and this remoteness raises the impediments to find collaborators.

From a "neoclassic" perspective, the partners in research collaborations aiming at spillover internalisation should be geographically distributed along the lines of unintended spillovers flows – i.e. large spillovers between two regions should result into many bilateral collaborative links. Collaborative links in turn reinforce spillovers of tacit knowledge: Spillovers cause collaboration causes spillovers.

The intensity of inter-regional knowledge flows is thought to depend on the potential for communication and for face-to-face contact between regions (Hussler 2003, p. 525), which in turn is related to the intensity of inter-regional personal interaction. The potential for communication is related to trade flows (Mélitz 2002, p. 24). Along these lines, international interaction among researchers is affected by factors similar to the determinants of trade, e.g. distance as a major impediment to face-to-face contact (Andersson/Persson 1993, p. 20). But even more than trade flows, knowledge flows depend on the interacting parties' ability to understand and absorb the respective concepts exchanged. Sharing the same communicative norms and standards facilitates understanding, particularly for tacit knowledge (Hussler 2003, p. 525). And differences in these norms are partly attributable to differences in language and culture, but as well to sector-composition disparities.

Moreover, interaction depends on specific characteristics of a country or region per se (so-called fixed effects). Those particularities may partly be attributed to country size: if domestic interaction is assumed to be less affected by transaction costs than cross-border interaction, the international links of a large country should hence be less important to it than to a smaller nation. This effect is familiar from economics of trade: The larger a country's GDP, the less the relative importance of external trade.

The following pages will be dedicated to empiric results on the potential of inter-regional interaction and spillovers, their determinants, and conclusions relevant to the aggregate scale.

2.2.1 Trade as a Spillover Catalyst

The notion of spillover flows following international trade patterns can be grounded on two basic concepts: First, the potential for communication is shaped by similar determinants as those of trade. Second, benefits from domestic R&D may arise to non-domestic producers if they import intermediate products. Coe/Helpman (1995, p. 862) explain the rationale for the latter: In the single-country case, they assume domestic factor productivity to depend on domestic cumulative R&D efforts, producing intermediate and final products. If R&D effort would exclusively be embodied in intermediate products and if all of these products were to be traded internationally, the cumulative R&D stock determining domestic productivity would be that of the entire world rather than only the domestic one.[26] Based on this reasoning, Coe/Helpman (1995, p. 869) estimate total factor productivity in dependence of domestic cumulative R&D and trade-weighted foreign cumulative R&D. The authors infer elasticities measuring the increase in an OECD country's (A's) domestic productivity with respect to a 1% increase in the R&D stock of another OECD member B. Unsurprisingly, the effect is the stronger, the larger B's R&D stock, and the more A trades with B. Coe/Helpman (1995) thus perceive the impact of B's R&D spillovers on A to be proportional to the importance of trade with B to the economy of A (corrected for differences in cumulative R&D).[27] Among EU

[26] Of course, there is a prerequisite: In order to allow for international R&D spillovers through traded goods, one has to assume that their producers do not price in all externality effects.

[27] Note: Coe/Helpman (1995) assume cumulative R&D stocks to be additive: i.e. the world stock of R&D is equal to the sum of national stocks. When accounting for the fact that substantial R&D efforts are dedicated to absorptive capability, the world stock of R&D (knowledge) should be less than the sum of its parts. In particular this would imply that a less R&D-intensive country would benefit more

members, for instance, elasticities are particularly large for a 1% increase in German R&D, and predominantly raise productivity in its small neighbouring states.

With respect to research collaboration, this implies that projects aiming at spillover internalisation should develop their links along the lines of the largest (absolute) spillovers, i.e. following trade flows adjusted for differences in cumulative R&D. Coe/Helpman (1995) hypothesise spillovers to follow the exchange of goods and services, thus spillovers should vary according to the mass of interaction between two regions or countries.

The degree of interaction also raises the potential for inter-regional personal contact of researchers. Informal personal contacts were already mentioned as a prime factor in the patterns of spillovers and collaborative links. And Mélitz (2002, p. 24) confirms the potential for communication and trade flows to be inter-related.

2.2.2 Informal Personal Contacts

As has already been discussed in section 2.1.6, informal social contacts and interaction are central to network formation and cooperation. Pohoryles (2002, pp. 334-335) analyses CIS data on projects in the fourth FP and highlights the importance of social factors: two thirds of researchers knew at least one partner personally or by reputation, more than two thirds knew one or more partner institutions through previous collaboration, and more than half of respondents previously even have worked for a project partner. Caloghirou/Constantelou/Vonortas (2001) list opportunities for the establishment of contacts: Previous experience once again constitutes the most important determinant, followed by conferences, respectively trade fairs. Informal personal contacts are equally perceived as the most essential factor for maintaining existing networks. Hussler (2003, p. 525) finds the frequency of face-to-face contact the most important factor for knowledge spillovers.

But how is inter-regional interaction between researchers determined? Beckmann (1993) and Andersson/Persson (1993) present models of academic collaboration (in particular co-authorship), the former on the micro- and the latter on an agglomerated scale. Beckmann (1993) finds collaboration probability to be largest among academics of similar “productivity” (Compare the symmetry argument, p. 36) and to decrease with the amount of travel time collaborating scientists need to invest. Partly based on Beckmann, Andersson/Persson (1993) investigate co-authorship between “creative” European regions and propose a gravity model, showing bilateral collaboration to depend over-proportionally on both regions

then estimated from an increase in advanced country B, while B would benefit less from an R&D increase in country A. This implies a bilateral spillover balance to diverge from the balance of embodied knowledge as assumed by Coe/Helpman (1995).

publication output and to be slightly impeded by geographic distance. Moreover, the authors find positive effects of bilateral adherence to a political or linguistic bloc. Conformingly, several other authors underline the importance of geographic and cultural/linguistic distance – empirical findings on this issue will be presented on the next pages.

2.2.3 Geographical Proximity

Geographic distance is the factor most frequently applied to describe spatial interaction patterns. Basically, it is perceived as a proxy for the cost of personal interaction, in particular in travelling. The amount of travel time needed to meet distant counterparts partly determines the frequency of face-to-face contact, and thus the probability of two random people to meet. If the intensity of personal networks is assumed to depend on the time dedicated to it by its adherents, distance thus decreases the probability for people to know each other.

The cost in money and time inflicted by geography also determines the patterns of trade, as recognised by most models of trade (compare Porojan 2000). Fischer/Tscherngell/Jansenbauer (2004) and Hussler (2003) (for European patent citations), as well as Andersson/Persson (1993) (for academic collaboration) analyse research interaction between European regions and nearly implicitly assume spatial distance as a major impediment to interaction between their geographically bounded nodes.

The latter two use a decreasing function of geographic distance in kilometres, whereas Hussler (2003) and trade economists (compare Porojan 2000) try to incorporate transaction costs with travel time, or the number of borders to cross, etc. In that context, it may be questioned whether spatial distance plays a role as important as in trade, for example. Modern communication devices have reduced nominal cross-border communication costs to a negligible minimum. For necessary face-to-face interaction, transport costs (particularly trip duration) definitely play a role. At the regional level, the costs for land-locked transport (cars, trains, ferries) increase with distance and will lead to less frequent contacts. However, at distances above a certain threshold air travel will adjust trip time and cost rather to traffic patterns (air hubs) than to geographic distance. Since we analyse collaboration on the European level, we assume many of face-to-face contacts associated with collaboration to be effectuated via air travel. The relationship of air travel cost to geography is considerably less strict than for other means of transport – it may be less time-consuming to go from Western Austria to Genoa than from Vienna to Paris. Therefore the decaying function of distance used in gravitational models of trade economics may either be subject to a relatively rapid decay, or break off above a certain level.

2.2.4 Language and Cultural Differences

Hussler (2003, p. 525-526) points out that particularly the transfer of (tacit) knowledge is facilitated if the interacting parties share the same set of norms and standards. She therefore reckons differences in cultural settings to slow down information and knowledge exchanges. Obviously, the ability to communicate depends not only on cultural concepts but also on language: Trade economists consider the potential of interaction and communication between regional populations as a major determinant for the geographic distribution of trade flows (Mélitz 2002). Complex and technical communication between persons is only feasible when both parties master a common language or dispose of adequate translation facility. Although translation may be adequate for some forms of trade, the complexity and dynamics associated with research cooperation lead us to omit it as a possible factor. We therefore focus on the communicative potential between persons, which is influenced by culture, micro-level conditions and language.

One might think that English has evolved already into a common standard for the interaction among researchers of different native tongues. However it is not yet a common standard for *all* of scientific interaction: publication statistics show that still a large part of scientific literature (particularly on the humanities and social sciences) is published in domestic tongues other than English (Sandelin/Sarafoglou 2003). The chances for spillovers arising from lesser-spoken tongues are much lower since non-speakers may not read these publications “by chance”. Moreover, we have to keep in mind that much research work is operated in a non-English language, even though the written research “output” may be published in English: Publishing in a certain idiom may not require the skills needed for oral interaction. In addition, the establishment of cooperation often follows prior informal personal contacts – and those are related to oral proficiency in the lingua franca, but even more in other common languages.

Mélitz (2002) explicitly studies the effects of language on international trade in a gravity model setting, and finds several language indicators to exert significant influence on the direction of trade flows. First, the more people in a country pair master a common idiom the higher bilateral trade flows. Second, linguistic diversity in a country promotes external trade (Ibid, p. 23). Moreover, he distinguishes between open-circuit idioms and spoken languages. The former term describes official, large languages spoken, while the latter one depicts the percentage of native speakers of the most important idioms. Mélitz (2000) concludes that “open-circuit” languages suffice for trade in homogenous goods, while “direct communication” pairs are needed for complex transactions. For our purpose, the effect on cooperation depends on whether English serves only as an “open-circuit” standard (for written communication) or enables partners to enact direct interaction.

We conclude that a measure depicting the extent of common language between countries as well as domestic linguistic diversity may have an impact on the choice of research partners. An assessment of proficiency per language (especially English) would possibly enhance this measure. It has to be noted that most European languages are confined to one or two member states. Moreover, scientists depict rather good knowledge of foreign idioms (European Commission 2001a). Therefore, the desired linguistic measure would have to account for foreign language knowledge.

Analogously to trade, foreign language knowledge is related to geographical proximity, but to cultural factors as well. Button et al. (1993) consider cultural relatedness an essential force in the direction of academic communication and base this conclusion on survey data. Complex or informal interaction is eased when communicating parties share, or understand a common cultural context. In analogy to Mélitz (2002), Rauch (1999) asserts that differentiated products and services are more likely to be traded among partners sharing a common cultural or historic background. We reckon that this may also hold for European research cooperation and state that cultural barriers may possibly either substitute language factors or provide supplementary information. Hussler (2003) tests differences in Hofstede's (1980) cultural dimensions for their impact on European patent citations: In relation to the other factors she includes (geographic distance, technological similarity, but not language), she finds cultural differences to play a minor role.

2.2.5 Technological Similarity and Sector Composition

Differing standards and norms as an interaction impediment do not only relate to language and culture, but also to sectors and research topics. When looking for knowledge exchange regarding a specific problem, one stands the most to gain from interaction with somebody facing the same type of problem. Conversely, several scientific disciplines and high-tech industries have developed a standard nomenclature, and forums such as journals or conferences to facilitate information exchanges on their research topic.

Caloghirou/Constantelou/Vonortas (2001) list conferences, respectively trade fairs as the second most important forum for establishing contact with future collaborators. On the international scale, we assume such conferences or fairs to be fairly specific, i.e. concentrating on a sub-topic appealing to a limited number of researchers and decision-makers. I.e. for highly specialised institutions it may be efficient to adhere to an institutionalised international network of face-to-face contact and information exchange. Hence the more two regions devote resources to similar research problems or similar technological sectors, the more researchers from both regions are likely to know each other in the forums, and the more cross-border contacts will be established.

Based on this rationale, Hussler (2003, p. 532) introduces “technological proximity”, a bilateral indicator denoting the correlation between two country’s patenting activities in 116 sectors. The more similar the patenting vectors, the more patent citations should occur between countries, so the presumption. Her empirical results strongly support the hypothesis. This is in line with the finding that spillovers and collaboration take place relatively more often between firms in the same sector than across sectors (inter alia Röller/Tombak/Siebert 1997).

2.2.6 Country-Specific Determinants of Collaboration

With respect to the gravitational approach applied in this study, we mentioned Fischer/Tscherngell/Jansenbauer (2004), Hussler (2003) and Andersson/Persson (1993), who estimate regional research interaction and find substantial mass effects: i.e. the interaction between two regions to depend positively on a multiplicative combination of the number of patents or publications in either region. Like other papers (inter alia Sharp 1998, Coe/Helpman 1995) they relate aggregate R&D figures for countries or regions to the resources devoted to R&D.

But apart from the sheer “mass” of R&D resources as a prime determinant of research collaboration, the interacting behaviour of regions and countries is frequently found to follow some factors unaccounted for in quantitative estimations. For this reason most of the author studying cross-country data introduce so-called dummy variables (country specific constants) to account for those fixed effects.[28]

Geuna (1998) and Caloghirou/Tsakanikas/Vonortas (2001, p. 157) study the impact of country fixed effects on university participation in European research cooperation. Geuna’s (1998) study focuses on pure university collaboration among EU-12 members: corrected for micro-effects presented in section 2.1, he finds departments from Germany, Spain, France and Italy to be significantly under-represented versus the UK, while institutions from most small countries (particularly Greece and Ireland) are relatively more involved in the FP.

Caloghirou/Tsakanikas/Vonortas (2001) analyse university-industry collaboration in the FP and confirm those disparities, albeit less pronounced: German and Italian universities participate significantly less than average, while again Greek and Irish (and Dutch) universities are considerably more involved in collaboration.

Marín/Siotis (2002, p. 27) and Hernán/Marín/Siotis (2003, p. 86) include country dummies into their estimation of a roughly similar sample of inter-firm RJVs in FP and Eureka. Both

[28] Inter alia: Geuna (1998), Caloghirou/Tsakanikas/Vonortas (2001), Marin/Siotis (2002), Fontana/Geuna/Matt (2003), Hernán/Marín/Siotis (2003)

find strongly significant, negative coefficients for Germany, France and Italy. Moreover, the former authors identify negative fixed effects in FP participation for the UK and positive effects for Belgium, Denmark, Ireland, Finland, and in particular Greece. But they also show Greece and Ireland to collaborate far less in the Eureka programme, which is not EU-funded.

The result renders clear why all the authors mentioned attribute the strong standing of Ireland and Greece, and especially of their universities to the “cohesion effect”: The EU Commission is known to favour projects with participants from less advanced “cohesion” countries. We regard the cohesion effect certainly to be of relevance – but the performance of the “periphery’s” academia may also be due to crowding-in/crowding-out effects: We assume an interest of the commission to distribute FP funds roughly “equally” among member states. Since large R&D-intensive companies are/were considerably harder to find in “cohesion” countries, research excellence may be concentrated at universities, which therefore constitute the “better” partners for cooperation. In the large and rich countries, corporate research partners may be easier to identify. Moreover, researchers in poorer countries with lower expenditure per researcher might be more inclined to search FP subsidies than their colleagues in richer states.

Geuna (1998) attributes the weak performance of universities from the large continental countries to the existence of large to the fact that these four countries dispose of large and excellent non-university research organisation, which may crowd out universities (e.g. the German Fraunhofer-Institut or the French Centre National de la Recherche Scientifique). We rather follow the argumentation of Hernán/Marín/Siotis (2003, p. 88), who conclude that large countries provide more opportunities for domestic cooperation – therefore, European cooperation is of lesser importance than to small member states. The more advanced researcher is in his topic the less likely he is to find a potential (symmetric) collaborator on that subject among a given population of researchers. Researchers in large EU countries are therefore more likely to find collaborators within the scope of domestic programmes, while their colleagues in smaller countries are more dependent on interaction with foreigners. The reader will recognise the similarity of the argument to the economics of international trade, where numerous studies have found larger countries to trade less intensely (Porojan 2000, p. 12).

Caloghirou/Tsakanikas/Vonortas (2001, p. 156) further indicate that Greece, France and the UK hold an over-proportional share of prime contractors in university-industry cooperation. While the importance of the latter two is mirrored in pure industrial collaborations (compare section 4.1.2), we attribute the Greek performance to the strong standing of its universities. Sharp (1998, p. 583) mentions that France and the UK are by far the most important recipients of EU exchange students with respect to population, while Germany hosts a relatively low number. We relate this to Pohoryles (2002, p. 334) finding that mobility patterns shape collaborative links, and conclude that France and UK have a more central position in academic research than equally research-intensive Germany.

2.3 Conclusions from Literature on Research Collaboration

The presented review of papers on research collaboration is far from being exhaustive. However, we hope to have got a hold of the key concepts, at least as far as they matter to this study. Obviously, the environmental factors enquired in section 2.2 are of more relevance to the agglomerated data set to be examined later on, while employing data on many of the micro-factors would be beyond the scope of this diploma thesis. Nevertheless, section 2.1 contributes important material on the mechanisms of research collaboration, which are predominantly situated on the micro scale. We will try to use several of the listed concepts with agglomerated data (firm size or research intensity, for instance). However we have to acknowledge that agglomeration of micro-effects runs the danger of common pitfalls: the combined effect of micro-factors may not be equal to that of the sum of their parts.

In order to facilitate the synopsis of the hypotheses and conclusions presented, we will summarise selected statements of the current section. Moreover we will quote the papers where this ideas stem from, the extent to which we judge the respective statements to be empirically validated and their relevance to our empirical enquiry in sections 6 and 7.

The selected statements *are not citations* but denote our personal interpretation of the respective authors' findings. Furthermore, the statements may not arise from the main purpose our conclusion of the corresponding paper, but may just represent a side-statement for its author(s). The classification of empirical content and relevance to this diploma thesis is solely based on our judgement and implies no statement for any other purpose whatsoever.

In sections 6 and 7 we attempt to model bilateral regional FP4 collaborative links with exogenous factors. In order to identify prospective explanatory variables, we will implicitly draw on the hypotheses laid out in Table 1. Among the most important rank the frequently cited impact of firm size (indexes 1 and 11) and of research intensity respectively absorptive capabilities (indexes 5 and 7). Since the data set under investigation consists of regionally agglomerated data, the environmental factors are of great importance, partly due to hard economic facts and partly because of the informal personal relationships they foster or impede (indexes 17 to 26).

Table 1: Selected literature statements of relevance to the empirical work

MASS FACTORS

Index	Selected Statement	References	An.	Emp.	Relev.	P.
1a)	Large firms collaborate more than SMEs	Hernán/Marín/Siotis (2000, p. 87), Miotti/Sachwald (2002, p. 1486), Caloghirou/Constantelou/Vonortas (2001), Marín/Siotis (2002)	☑	●	●	17, 36
1b)	The propensity to collaborate depends positively on market share, i.e. the relative firm size with respect to the market.	Miotti/Sachwald (2003, p. 1490), Link/Bauer(1989)	☐	◕	◔	37
1c)	The number of a university department's staff members is positively related to its number of collaborative links (but has no impact on the decision whether to participate in general)	Geuna (1998, p. 686)	☐	◕	◔	36
1d)	Large firms are more inclined to UI collaboration	Miotti/Sachwald (2003, p. 1491), Fontana/Geuna/Matt (2003, p. 19)	☐	●	◔	38
2a)	Firms participate in RJVs for cost-sharing (particularly large firms)	Röller/Tombak/Siebert (1997, pp. 4-8)	☑	◕	◑	22
2b)	RJVs are formed for reduction of time-to-market	Caloghirou/Ioannides/Vonortas (2003, p. 553)	☐	◔	◔	23
3a)	Firms participate in RJVs for spillover internalisation (particularly large firms)	Röller/Tombak/Siebert (1997, pp. 4-8), Kaiser (2002)	☐	◕	◑	23
3b)	The internalisation of spillover effects leads to RJVs with a large number of participants	Röller/Tombak/Siebert (1997, p. 9) Hernán/Marín/Siotis (2003, p. 85)	☑	◔	◔	23
3c)	The need for the definition of standards fosters large RJVs	Luukkonen (2002, p. 449-450)	☐	◔	○	25
4a)	Large firms are more interested in pre-competitive RJVs, SMEs more in marketable applications	Luukkonen (2002, p. 441)	☐	◑	◑	32
4b)	Low-tech firms perceive RJVs as an opportunity to perform basic research	Luukkonen (2002, p. 442)	☐	◔	◔	35
5a)	RJV participation depends positively on absorptive capacity or "research intensity" of a firm	Link/Bauer (1989), Caloghirou/Constantelou/Vonortas (2001), Miotti/Sachwald (2002, p. 1490), Fontana/Geuna/Matt (2003, p. 19), Kaiser (2002)	☑	●	◑	27-29
5b)	High-tech firms cooperate more	Miotti/Sachwald (2003, p. 1486)	☐	◕	◑	28

5c)	Scientific excellence (publications per researcher) of universities promotes FP collaboration	Geuna (1998, p. 683)	☐	◕	◑	29
6a)	R&D collaboration is performed to enhance absorptive capacity and tacit knowledge, especially with universities	Caloghirou/Tsakanikas/Vonortas (2001, p. 159), Cohen et al. (1997)	☐	◑	○	30
6b)	Scientific advances as such and reputation enhancement are major collaboration incentives for universities	Caloghirou/Constantelou/Vonortas (2001), Hakala/Kutinlahti/Kaukonen (2002, p. 370)	☐	◑	○	29
7a)	Firms at “technological frontier” collaborate more with foreign partners	Miotti/Sachwald (2003, p. 1492)	☐	◑	◑	30
7b)	„Hard“ sciences are more international: technical universities collaborate more at international level	Sandelin/Sarafoglou (2003, p. 4), Geuna (1998, p. 683), Hakala/Kutinlahti/Kaukonen (2002, p. 358)	☐	◑	◷	31
7c)	Intense foreign competition increases likelihood of RJV formation and the choice of foreign collaborators	Link/Paton/Siegel (2002, p. 1463)	☐	◕	◷	31
8a)	Access to public subsidies increases RJV formation (no crowding-out)	Veugelers (1997, p. 311), Miotti/Sachwald (2003, p. 1489)	☐	◕	○	39
8b)	Funding reasons are a main incentive for universities’ international collaboration	Hakala/Kutinlahti/Kaukonen (2002, p. 371), Pohoryles (2002, p. 334)	☐	◕	○	40
8c)	Importance of framework programme is propelled by the degree of a public organisation’s dependence on competitive funding	Geuna (1998, p. 686), Van der Meulen (2002)	☐	◑	◷	40
9)	A business cycle downturn increases the need for cost-sharing and thus for RJVs	Link/Paton/Siegel (2002, p. 1463)	☐	◔	○	40
10)	RJV participation probability increases when a firm has participated in RJVs before	Hernán/Marín/Siotis (2003, pp. 84-85), Caloghirou/Constantelou/Vonortas (2001), Fontana/Geuna/Matt (2003, p. 10), Marín/Siotis (2002)	☐	●	◔	28

DISTANCE FACTORS

	Selected Statement	References	An.	Emp.	Relev.	P.
11a)	Symmetries in firm size promote bilateral collaboration	Röller/Tombak/Siebert (1997, pp. 5-11)	☑	◕	●	19
11b)	Symmetries in universities’ reputation raises prospects for bilateral collaboration	Geuna (1998, p. 679)	☐	◔	◷	37
12a)	Complementarities in firms’ R&D resources promote bilateral collaboration	Katsoulacos/Ulph (1998), Caloghirou/Ioannides/Vonortas (2003, p. 556)	☑	○	◷	21
12b)	Complementary final products provide no incentive for firms to enter RJV; in case of symmetric products spillovers and cost-sharing foster collaboration, but only up to a certain degree of symmetry	Navaretti et al. (2002, pp. 36-37)	☑	◑	◷	19

12c)	Partners in vertical production are more inclined towards collaboration	Röller/Tombak/Siebert (1997), Kaiser (2002)	☐	◕	◔	25
12d)	Collaboration with vertical production partners is more likely for small and for low-tech firms → small effect on international RJVs	Luukkonen (2002, p. 443), Miotti/Sachwald (2003, p. 1491)	☐	◑	◔	26
12e)	Adherence to a corporate group fosters collaboration, particularly among group entities	Miotti/Sachwald (2002, p. 1486, pp. 1490-1495)	☐	◕	◔	26
13)	Informal personal contacts foster formal bilateral collaboration	Caloghirou/Constantelou/Vonortas (2001), Pohoryles (2002)	☐	●	◕	38
14)	Firms employ their resources in many RJV participations for risk diversification	Pyka/Windrum (2001, p. 10)	☑	○	◔	22
15a)	IPR issues hinder University-Industry (UI) collaboration	Hall/Link/Scott (2001, p. 89)	☐	◕	◔	32-36
15b)	UI collaboration is the more likely, the more the research results are uncertain and the less they are appropriable; i.e. UI collaboration is inclined towards basic research	Hall/Link/Scott (2001, p.94), Baldwin/Link (1998)	☐	◕	◑	32-36
15c)	Universities participate more intensely in RJVs with a large number of participants and a long duration: i.e. basic research	Caloghirou/Tsaknikas/Vonortas (2001, p. 157)	☐	◕	◔	32-36
15d)	Firms relying heavily on patents (biotech, pharmaceutics, chemicals) are more inclined towards UI collaboration	Caloghirou/Constantelou/Vonortas (2001, p. 14), Miotti/Sachwald (2002, p. 1491), Luukkonen (2002), Fontana/Geuna/Matt (2003, p. 20)	☐	●	◔	31

ENVIRONMENTAL AND MISCELLANEOUS FACTORS

	Selected Statement	References	An.	Emp.	Relev.	P.
16)	Market concentration eases identification of partners and spillover internalisation, thus leads to more collaboration	Hernán/Marín/Siotis (2003, p. 84)	☐	◔	◔	40
17)	Large countries (Germany and Italy in particular) exhibit less FP collaboration intensity then small countries (especially Greek, Irish and Dutch universities)	Geuna (1998, pp. 684-685), Caloghirou/Tsaknikas/Vonortas (2001), Hernán/Marín/Siotis (2003, pp. 86-88)	☐	●	●	47
18a)	Trade is related to the potential for communication	Mélitz (2002, p. 24)	☐	◑	◑	41
18b)	Spillovers are related to the potential for face-to-face contact	Hussler (2003, p. 525)	☐	◑	◑	41
19)	Spillovers are the higher, the more important a trade partner is to the recipient	Coe/Helpman (1995, p. 873)		●	◕	42

20a)	International academic interaction depends negatively on distance and follows roughly a gravity model pattern	Andersson/Persson (1993, p. 20), Beckmann (1993, p. 8)	☑	◐	●	43
20b)	The “number” of spillovers depends on the “mass” of counterparts	Hussler (2003, p. 534), Coe/Helpman (1995), Fischer/Scherling/Jansenberger (2004)	☐	●	●	47
21)	Geographic distance decreases knowledge spillovers	Fischer/Scherling/Jansenberger (2004, pp. 7-10), Hussler (2003, p. 533)	☑	◕	●	44
22)	Cultural differences matter for the degree of knowledge spillovers	Hussler (2003, pp. 529-536)	☐	◐	●	45
23)	Bilateral trade depends positively on the relative share of people speaking a common language	Mélitz (2002)	☐	◕	●	45
24)	Mobility patterns of researchers and students shape collaboration distribution	Pohoryles (2002, p. 334)	☐	◐	◐	49
25)	Similarities in technological structure foster bilateral spillovers	Hussler (2003, p. 534)	☐	●	●	46
26)	Remoteness hinders trade	Porojan (2000, p. 4)	☐	●	●	41

Index: Provides grouping of selected statements adhering to a common nexus.
Selected Statement: The author's interpretations of conclusions by studies provided in “References”.
References: Study or studies supporting the selected argument theoretically and/or empirically.
An.: Indicates whether conclusion derives from an analytical model. If several references are provided, not all of them may be based on such a model.
Emp.: The extent to which we judge the respective argument empirically validated by one or several of the authors listed in “References”.
Relev.: The extent to which we judge the respective argument being of relevance to data selection, empirical analysis and interpretation carried out in sections 5-7.
P.: Refers to the page in this document where the selected statement is presented.

3 THE EUROPEAN FRAMEWORK PROGRAMME – AN INSTITUTIONAL OVERVIEW

This chapter is dedicated to an institutional specification of our research subject, the European Framework Programme (FP) collaboration schemes. The first section will briefly summarise the FP's historic evolution, its structure and the criteria for subsidy awards. The second section will specify the FP's institutional setting in the context of other research cooperation programmes, particularly at the European level. The final part will sketch the economic and political reasons for public promotion of research cooperation, and for the creation of FP in particular.

Since the empirical part of this study focuses on data from the Fourth Framework Programme (FP4), the institutional properties presented will concentrate on the FP4 in particular.

3.1.1 The 4th Framework Programme – Evolution & Objectives

Support of R&D has evolved into one of the four large items in the European Union budget, of which the framework programmes constitute the by far most important part.

The issue of supra-national research policy coordination draws its origins back to the treaties of Paris and Rome, but judged by the proclaimed aims, the various initiatives can largely be characterised as a failure (Banchoff 2002, p. 4). Throughout the 1960s and 1970s, the need for sharing the cost of large projects was mainly satisfied through inter-governmental rather than supra-national cooperation (as, for instance, in CERN, ESA, etc.). The most noteworthy collaboration scheme from this era was the mainly national-level dominated COST programme (see p. 62).

Since the 1960s the sentiment of a European "technology gap" vs. the US (and later Japan) provided the reason for calls to consolidate national European research policies. This situation intensified in the early 1980s, when the perceived Japanese technological supremacy was attributed to innovation policy administered by the Japanese ministry of international trade and industry (MITI). In 1981, a "round table" constituted of the commission and 12 large European electronics firms launched the pilot phase of the ESPRIT programme.[29] Imitating MITI-initiated business R&D, the supra-national scheme aimed to foster intra-European corporate collaboration, stressing the “pre-competitive” character of cooperation. ESPRIT became the archetype for and the nucleus of the later FP, being joined by its sibling RACE for communication technologies in 1982 (Georghiou 2001, pp. 892-893). Several similar schemes followed focusing on other R&D topics. The increasing number of R&D collaboration schemes required administrative effort by the Council, which in turn led to the consolidation of those into the framework programme. At the time the so-called “Reisenhuber criteria” (named after the German research minister) were formed to delineate the FP’s objectives and position in the European space of RTD policy. Quintessentially, they argued for supra-national support when:

“[…] the scale or cost of co-operation was beyond that affordable by a single country, where complementarily in national work could achieve results for the whole Community, and where research contributes to development of the common market, laws and standards, or to the unification of European science and technology.” (Georghiou 2001, p. 893)

This formulation draws on the principle of subsidiary, but also leaves space for the frequently cited aim to strengthen the technological base of the EC. Moreover the criteria do not refer to the “pre-competitive” character demanded in the implementation of ESPRIT. In practice, the FP’s already numerous sub-programmes provided support from basic research up to near-market R&D (Georghiou 2001, p. 894). In addition to the “competitiveness-oriented” Reisenhuber criteria, the objective of “cohesion” was introduced with the Single European Act 1987. Despite the increasing importance of the FP, numerous European RTD activities were set outside the FP’s range. From 1988 to 1990, EC expenditures on R&D increased from 2.6% of the budget to 4%, to equal parts due to expenditure rises in FP and in other schemes.

Moreover, the late 1980s saw the rise of Eureka, a collaborative RTD policy targeting nearer-to-market industrial RJVs and supervised on an intergovernmental basis as opposed to the Commission’s dominance of FP (see p. 63).

The Articles 130 f-130 p of the Maastricht Treaty[30] 1992 broadened the scope of European RTD policy and, apart from strengthening the technology base, explicitly allowed involvement into areas of Community interest such as energy or transport (Georghiou 2001, p. 894).

[29] ESPRIT: European Strategic Programme for Research and Development in Information Technologies

[30] European Union (1992). Due to the Amsterdam treaty’s re-numbering of articles, the relevant passages in the current EU treaty now are concentrated in Title XVIII (Articles 163-173). (European Union 1997).

Article 130 i names the "multi-annual framework programmes" (European Union 1992) as the central European RTD policy tool. For the first time, Article 130 h assigned the Community the competence to promote coordination between supra-national and member state RTD policies (Sharp 1998, p. 570). However, efforts by the late Delors commission to ameliorate coordination were shelved in the aftermath of the pan-European recession in 1993. Conversely, the FP4 (scheduled for the period 1994-1998) received a near doubling of funds and broadened its involvement (see Table A.2 in the appendix). However, the major part of the budget increase stems from the inclusion of schemes formerly outside the FP: Apart from the formerly out-of-balance "accompanying measures", the innovation scheme SPRINT and the energy programme THERMIE were incorporated into the FP4 budget.[31] The FP4 now comprised the entirety of European Union expenditure on RTD, stabilising at a share of about 4% of the total EU budget. Despite the extended range of FP instruments, the collaborative schemes still provided for the by far most important part of the FP4's cost (79 % of total[32]).

The subsequent FP5 once again widened the range of sub-programmes and increased its budget by about one-tenth vs. its predecessor. The Amsterdam treaty 1997 introduced social objectives and reduced the quorum for FP-related Council decisions from unanimity to a qualified majority (Georghiou 2001, p. 894). But in reality the strategic orientation and the practical implementation did not change vs. the FP4.

The 6th programme implemented from 2002 on, in contrast, was incorporated into the broader aim of establishing a European Research Area (ERA). In accordance with the Lisbon catalogue (Banchoff 2002, p. 2), the ERA aims for a truly integrated "market" for knowledge generation and innovation and puts European RTD instruments in the spectrum between research and innovation policy.

Although the 5th and particularly the 6th framework programmes provide interesting research topics, the data set we focus on in this study draws on results from the FP4. Therefore we will henceforth discuss FP policy only up to 1998 (the final year of FP4).

3.1.2 Structure of the Fourth Framework Programme

As has been mentioned, the FP4 was the first to include virtually all of EU RTD policy. The entire budget approved initially amounted to 12 300 m€, later expanded by 800 m€ to

[31] SPRINT: Strategic Programme for Innovation and Technology Transfer (SPRINT late became the "third activity" of the FP4); THERMIE: Technologies Européennes pour la Maîtrise de l'Energie

[32] "Indirect Actions" adhering to "Activity 1" as proportion of the total FP4 budget over the entire period 1994-1998. (European Commission 1998, p.89)

accommodate the three member states entering the Union in 1995. Together with a final complementary financing of 115 m€, the total budget thus covered 13 215 m€.

The structuring of the FP4 budget may be organised along four dimensions:

Table 2: Budget for the Fourth European Framework Programme[33]

	Fourth Framework Programme Decisions 1110/94/EC, 616/96/EC, 2535/97/EC			Euratom Framework Decisions 94/268, 96/253/Euratom				
	Indirect actions	JRC	Support for DGs	Indirect actions	JRC	TOTAL	% OF TOTAL	Responsible
FIRST ACTIVITY								
Information and communication technologies	3646.0	11.5	10.5			3668.0	27.8%	
1. Telematics applications	913.0					913.0	6.9%	DG XIII
2. Communications technologies	671.0					671.0	5.1%	DG XIII
3. Information technologies	2062.0	11.5	10.5			2084.0	15.8%	DG III
Industrial technologies	1921.0	208.5	10.5			2140.0	16.2%	
4. Industrial and materials technologies	1737.0	96.0				1833.0	13.9%	DG XII
5. Standards, measurement and testing	184.0	112.5	10.5			307.0	2.3%	DG XII
Environment	816.5	313.0	27.5			1157.0	8.8%	
6. Environment and climate	573.5	313.0	27.5			914.0	6.9%	DG XII
7. Marine science and technology	243.0					243.0	1.8%	DG XII
Life sciences and technologies	1627.5	50.0	31.5			1709.0	12.9%	
8. Biotechnology	595.5					595.5	4.5%	DG XII
9. Biomedicine and health	374.0					374.0	2.8%	DG XII
10. Agriculture and fisheries	658.0	50.0	31.5			739.5	5.6%	DG XII
Energy	1039.0	21.0	16.0	1016.5	319.5	2412.0	18.3%	
11. Non-nuclear energy	1039.0	21.0	16.0			1076.0	8.1%	DG XII
12. Nuclear fission safety				170.5	270.5	441.0	3.3%	DG XVII
13. Controlled thermonuclear fusion				846.0	49.0	895.0	6.8%	DG XVII
14. Transport	263.0					263.0	2.0%	DG VII
15. Targeted socio-economic research	112.0	35.0				147.0	1.1%	DG XII
SECOND ACTIVITY								
Cooperation with third countries and international organisations	575.0					575.0	4.4%	DG XII
THIRD ACTIVITY								
Dissemination and utilisation of results	312.0		40.0			352.0	2.7%	DG XIII
FOURTH ACTIVITY								
Stimulation of the training and mobility of researchers	792.0					792.0	6.0%	DG XII
TOTAL	11104.0	639.0	136.0	1016.5	319.5	13215	100.0%	

33 Adapted from European Commission (1998), p.89. Source for responsible DG: CORDIS (2004)

Concerning the relevant Council decisions, 90% of RTD funds stem from the European Communities (EC) budget, with the relevant decisions grounding on the Treaty of the European Union (1992). The remaining 10% stem from the means of Euratom, which is a separate legal entity and provided for about half of the nuclear fusion and fission research budget.

Moreover, funding is divided among four “activities”: Activity 1 encompasses the RTD schemes for with the FPs initially were set up, and accounts for 87% of the total budget. The second, third and fourth Activities (so-called “horizontal” activities) are formerly independent EC programmes included within the FP scheme from 1994 on. The objective of Activity 2 is to provide the funds necessary to foster extra-EU FP collaboration, while Activity 3 incorporated the formerly independent SPRINT schemes to promote the “application” of newly acquired technologies and advance SME innovation. Finally, Activity 4 aims at the social integration of European researchers and funds networking, fellowships and prizes. The database investigated in this paper draws solely on indirect actions, of which the large majority is carried out under Activity 1. Moreover in the consolidated and regionalised data sample Activity 1 is even more dominant; therefore we omit further description of the remaining Activities.

In terms of recipients, Activity 1 divides its budget between two main types of action: “Direct Actions” refers to the funding of the European Union’s “Joint Research Centre” JRC. The JRC is the EU equivalent to the large national research organisations (albeit the JRC disposes only of 2,000 staff) and its seven institutes carry out research of “common interest of the Member states” (JRC 2004), mainly in nuclear energy. However, the JRC never gained the position it was intended for and remained of lesser importance, evidenced in its Activity 1 budget share of 4%.

With 10,441 m€, or 79% of the EU RTD budget, Activity 1’s “Indirect Actions” represent by far the largest chunk of European research funding. This scheme supports the collaborative research projects commonly identified with the FP. The collaboration programme divides into 15 “key actions”, i.e. broad topics to which the single projects adhere. The thematic sub-division developed out of the separate pilot initiatives (like ESPRIT or RACE) and is due to the fact that several key actions are administered by Directorate-Generals (DG) other than the one for research (DG XII). The distribution of funding evinces the focus on “applied” natural sciences, which suggests an important share of participating organisations in the corresponding fields (for instance technical universities, electronics groups, etc.). Moreover, the key actions represent R&D topics where collaboration is affected differently by motivations such as economies of scale, etc. Accordingly, project sizes (in terms of budget and of the number of participants) vary widely over key actions. Dachs/Roediger-Schluga

(2003), for instance, compare communications and agriculture key actions and find the characteristics of participants to differ considerably: While agriculture is relatively dependent on public research organisations, large firms dominate the communications section.

In this study, we will analyse collaboration data agglomerated over all 15 key actions and thus forego their differing characteristics. However, the reader may bear in mind that those differences may shape regional collaboration patterns, which this paper leaves unaccounted for.

3.1.3 Criteria for Project Acceptance

Equally to thematic orientation, the eligibility criteria for FP admittance affect the distribution of collaborative arrangements:

Submitting a project proposal to FP institutions is generally deemed to be a relatively bureaucratic procedure (Cornet 1999, p. 47). Applicants take care to conform to tough criteria, particularly since the majority of proposed projects are not deemed eligible for FP funding (for instance, only 29% of received proposals were selected for funding in 1997).[34]

The screening procedure each proposal has to undergo starts with a "call for proposals", issued annually or biannually by the respective organising body (e.g. DG XII). Proposals have to be submitted until the specified deadline, and are subject to a pre-check whether administrative criteria are met. Most notably, those criteria set the minimal number of participating member states and request the signatures of each party to the project. The about 95% of proposals passing this check are then classified as "eligible" and forwarded to more profound examination.

A panel of at least three independent experts in the corresponding fields reviews each eligible proposal. As an outstanding feature, the documents for expert review are anonymous, i.e. the project partner's names are not revealed to the experts before they have set marks for the evaluation criteria set in the "call for proposals".

In the FP4, those evaluation criteria were set separately for each key action, though all of them included the objectives outlined in the Maastricht treaty (European Union 1992, Art. 130 f - 130 p). (For the subsequent FP5, evaluation criteria were consolidated and expanded to social objectives.) As a representative example, ESPRIT (1997), requested the following conditions, structured in line with the FP5 objectives outlined in European Commission (2001b, pp. 13-14).

[34] Figure based on data in European Commission (1998, pp. 77-80)

- Scientific/technological quality and innovation
    - The proposal has to be up to the key action's topic and objective in terms of quality and thematic intent.[35]
    - The degree of innovation, and its state-of-the-art methodologies
    - A clear definition of what is why to be investigated from whom, given in a "Technical Summary".
- Resources, partnership and management
    - A detailed work plan, a clear assignment of partners' roles and an estimate of total project cost.
    - Appropriateness of the resources (particularly manpower) assigned in the project plan
    - Specified project objectives and their measurement
- Industrial Relevance
    - Clear relatedness to existing or anticipated market demand
    - A view of the project's positioning in the relevant market segments
    - An outline of potential impact on industry sectors and working conditions
    - Fostering innovations in goods, services or processes with impact on society
- European Dimension
    - The project contains at least two participants from at least two countries
    - "European added value of the consortium", i.e. carrying out the project at the European level has greater impact than the sum of comparable national projects; respectively a critical mass has to be attained for success.

While "Industrial Relevance" differs in scope according to the respective key actions, the remaining evaluation criteria are roughly the same for all key actions. Noticeably, in neither of the both documents cited above, the aim of "cohesion" is explicitly mentioned.

The experts individually mark projects according to the evaluation criteria and consolidate their assessment in a final meeting. Based on experts' recommendation, Commission officials rank the projects on a final list. With respect to the budget earmarked for the call, the top projects are selected for funding in descending order of priority. The whole procedure is subject to monitoring by independent referees.

---

[35] The ESPRIT programme's objective, for instance, was defined as: "[…] *to provide and demonstrate the technological building blocks for information society applications and for application in industry to strengthen the competitiveness of all EU industry. The tasks are described in the 1997 Esprit Work programme."* ESPRIT (1997)

The excursion on evaluation criteria allows several conclusions on the distribution of FP participants: First, it has been widely stated that the tough conditions of the selection process discriminate against SMEs. Therefore, we think that a disproportionate number of firms are selected that dispose of the resources to master the procedure. Second, the criteria are demanding in terms of scientific capabilities, focusing on organisations at the leading edge of their scientific field. Moreover, most of the key actions presented seem to focus on fields associated with significant economies of scale (with economies of scale being themselves a justification for the "European dimension"). Together with the advanced capabilities (and thus experience) demanded, this once again strikes the case for big entities already involved in large-scale research.

Although not explicitly demanded by the FP, the issue of cohesion matters as well to the project selection process, as is stated in Sharp (1998, p. 586):

"It is well known that the Commission look more fore favourably on consortia that include cohesion partners and may ask groupings to widen their membership to this effect."

The term "cohesion partners" relates to project partners from cohesion countries or regions (i.e. Spain, Greece, Ireland, Southern Italy, Portugal). Thus bidding groups are inclined to include at least one partner from the cohesion countries, even if its contribution would not have justified its inclusion in the first place.

3.1.4 FP in the European Technology and Innovation Policy Spectrum

Based on the criteria laid out in the Maastricht treaty, the FP is designed to foster pre-competitive interdisciplinary collaboration. The scheme was founded in an era when the distinction between technology and innovation policy just began to emerge. The latter clearly plays a relatively minor role, with the main programme on innovation (Activity 3) only accounting for 2.7% of the total research budget. The lack of focus on innovation has been widely recognised, even by the Commission (Georghiou 2001, p. 895). However, it may be argued that innovation policy is better conducted at the regional level, in line with the subsidiary principle.

In general, the FP provides less than 5% of total RTD expenditure in the EU-15, suggesting that its impact on shaping technology and innovation is relatively minor compared to the effect of national policies.[36] Given these comparably less significant figures, a focus on the FP's "core competence", the European cooperation, seems desirable. Thus the support of collaboration as such appears justifiable. In addition, many voices have called for supra-national coordination of member states policy for reasons presented on p. 67. As to be judged from the history until now, the EU clearly failed on that objective. Despite numerous reports on the current RTD situation and policies in member states, and repeated initiatives to "Europeanise" technology strategy, the EU still exerts virtually no influence on the design of national RTD policies (Banchoff 2002, pp. 7-10).

The FP's role thus appears as supplementing national programmes. This is particular true since most member states have their own RTD collaboration schemes in place. The quality and available funds from these national programmes may considerably affect the decision whether to cooperate nationally or cross-borders: I.e. apart from the effect that organisations in larger/richer countries find more eligible partners within their borders, the more intense public collaboration support in these states may bias them further towards national project partners. In addition, a differing strategic orientation of RTD policies across countries may affect composition of national FP participants. If, for instance, national RTD funds foster nearer-to-market projects while lacking support for basic research, domestic organisations focusing on the latter may over-proportionally participate in the FP.

With respect to international collaboration, the FP is also subject to competition from two other European R&D cooperation programmes: Eureka & COST, to be outlined below.

COST (European co-operation in the field of scientific and technical research) is the lesser known of both, having been established in 1971 by 19 countries. It provides for co-ordination of nationally based activities, with the organisation itself only funding coordination costs. Each member state has the right to propose four-year-long "actions" between at least five different project partners, with research being funded out of national budgets. In total, national funding of COST adds up to estimated 2,000 m€ annually (COST 2004).

The main feature of COST is its flexibility, since it does not subject projects to pre-defined topics. Moreover it is characterised as bottom-up, with initiatives and scope of activities determined by the research community (Marín/Siotis 2002, p. 21 and Georghiou 2001, p. 897). These characteristics are mainly cited for describing its success throughout the 1990s (where the number of actions rose from less than 50 in 1990 to roughly 150 in 1999). Conforming to flexibility, COST exhibits a wide range of research topics, and interestingly,

---

[36] However it has to be noted that the FP is of far larger importance to smaller and poorer member states such as Portugal or Greece. Moreover, such countries receive important RTD support from the structural funds (Sharp 1998, pp. 580-581).

the participant distributions shows more than three quarters of partners stemming from countries other than the “big five”. Due to the wide thematic distribution of COST projects, it appears to be less focused than the FP and Eureka; therefore we deem it rather as “hands-off” and a comparatively less active policy tool.

Dissatisfaction with European Union research support in the early 1980s led to the establishment of ***Eureka***[37] as a nationally controlled “industry-led” collaboration scheme. Eureka was intended to be nearer-to-market than the FP and to focus even more on high-tech industrial applications. Eureka as such does not dispose of any noteworthy funds. Instead, it “labels” accepted projects with its name. This “label” then renders the project eligible for national funding in the partners’ respective countries.

Several characteristics of Eureka differ versus FP in a manner perceived positively by project partners: First, it allows projects to be changed *en route*, second it does not place constraints on partner selection (i.e. no “cohesion” intentions), third it provides a clear legal framework with respect to competition law as well as to IPR and confidentiality protection. Finally, the administrative burden required for proposal acceptance is perceived as fairly low: Georghiou (2001, p. 895) characterises its style as

“[…] ’bottom-up’ and (relatively) non-bureaucratic with a very small secretariat.”

These characteristics led to overwhelmingly positive evaluation by project participants during the 1990s. Nonetheless, Eureka declined in importance between 1993 and 1998 (Georghiou 2001, p. 896). During the time of the FP4 (1994-1998) the total funds committed by private and public parties totalled between 2 b€ and 3 b€ (Eureka 2003).

Positioning itself nearer to marketable research than the FP, Eureka comprises a much larger share of firms among its project partners than does the FP. Large firms constitute by far the most important type of participants, even if their share declined against a rising number of participating SMEs. The near-to-market focus also shows up in the thematic distribution of projects, with “information technology” (IT) comprising the major part of projects and funds. Drawing on project numbers, we conclude that Eureka’s IT projects had and have a serious impact on participation in the FP’s information and communication action lines. Firms may prefer to effectuate R&D directly leading to marketable results within Eureka, while they may choose FP for “basic” research.[38]

---

[37] Eureka is not (anymore) an acronym. Its name stems from the original intention to create a European research coordination agency. (Georghiou 2001, p. 891)

[38] This hypothesis could mean that firms do their basic research with the FP and than advance the innovation stage to Eureka. However, this is not confirmed by empirical evidence, since traffic between

The geographic pattern of project distribution reflects the fact that organisations from virtually every European country may participate in Eureka schemes. The shares of EFTA and CEEC countries are therefore higher than in the FP. On the one hand, we may conclude that large high-tech firms from leading edge countries may over-proportionally participate in Eureka as opposed to the FP – but on the other hand very large firms may have a stronger interest in basic research than their smaller counterparts more behind the technological frontier. Therefore the characteristics outlined do not allow hypothesising on Eureka's effect on FP composition. However, comparing the geographic distribution of Eureka and FP in some thematic fields could provide a hint on the "cohesion effect" in FP partner choice (which we omit for lack of data).

3.1.5 Economic Rationale for and against Public Intervention in Research

The rationale for the FP, at least as it has been stated its legal text, remains somewhat unclear from an economist's perspective. But the vast majority of economic literature concerned with the topic has argued for support to research collaboration, in particular on the European level. Nevertheless, it is still disputed whether the FP constitutes the right answer to the commonly identified need for public intervention.

Since the late 1950s, the main part of economic literature on research has concluded that pure self-organisation of the R&D sector leads to under-investment, since knowledge production falls prey to many types of market failure, to be corrected for by public policy. Hauknes/Nordgren (1999) list the classic arguments in favour of public intervention and categorise them as descendents of the "Arrow/Nelson rationale". Nelson (1959) and Arrow (1962) ground their considerations on the economic theory of the firm. Basically, technological progress is perceived as embodied in knowledge, which has several distinct characteristics: knowledge is generic, codified, immediately accessible and directly productive. Thus if knowledge is produced once, it can be applied in the production process of the entire firm population at no marginal cost. From this characterisation of knowledge, three market-failure properties of knowledge may be inferred: First, knowledge is a quasi-public good, thus difficult to appropriate. Second, knowledge generation is subject to considerable uncertainty and, third, indivisibilities. (Compare Hauknes/Nordgren 1999)

Subsequently, literature has elaborated the implications of these market failures. Concerning international research collaboration, substantial economies of scale and the positive

---

both programmes is deemed as rather low (Georghiou 2001, p. 898). It is rather likely that firms divide into distinct groups, with the more basic research-oriented being relatively more involved in the FP.

externalities with risk sharing raise the incentives to *decrease research duplication*. From the economics of networks perspective, similar arguments strike the case for uniform, global *standards*. The imperfect appropriability of returns from knowledge generation goes hand in hand with *spillovers*, which may inflict externalities across borders. The assumed low copy costs imply that basic research should be rendered publicly available to allow for rapid diffusion and subsequent application in production. This enforcement of knowledge's public-good character increases the need for public support in its generation. But low copy costs argue as well for the optimality of *rapid cross-border diffusion*. (Hauknes/Nordgren 1999, and Cornet 1999).

Later research asserted the neoclassical framework to put too much emphasis on the public good character of knowledge and to underrate the personal element (Sharp 1998, p. 572). Absorptive capabilities rank among the most frequently cited prerequisite for the incorporation of external knowledge. This term circumscribes the ability to understand codified information and put it into practice. Part of this ability is grounded on non-codified "tacit knowledge", which in turn can only proliferate through personal contact (Caloghirou/Constantelou/Vonortas 2001, pp. viii-x). Research collaboration can provide many of the channels enabling such flows of tacit as well as codified knowledge (Ibid., p. xxiii).

Hauknes/Nordgren (1999) list several failures not explicable in the neoclassic framework: *Institutional failures* are due to the malfunctioning of soft institutions (laws, IPR, etc.) and hard institutions (i.e. organisations). A wrong calibration of the latter according to their functions (basic research and training for public organisations, innovation for private firms, etc.) may result in a mix producing sub-optimal output. *Network failures* relate to inadequate linkage between researching institutions: A network should improve the organisations' resource base and degree of freedom in a dynamic environment, lower risk and increase coordination. Policy's task in that respect is to support institutions and networks in order to achieve the "right" pace of variety generation and selection. Variety generation is defined as the working-out of new ideas, and selection is the process of converting this impetus into production. Network links may improve the prospects for variety generation itself as well as for selection.

Out of theoretic arguments, mainly three intertwined aspects strike the case for public support of research collaboration in general: spillovers, learning and the network-variety context. For economic purposes, learning is mostly modelled as a part of spillover absorption: As a main difference to classic spillover models, an absorber has to dispose of resources in order to integrate external knowledge. Held in the broader perspective of research networks, spillovers are just a (unidirectional) side effect of communication between persons in institutions. Variety (i.e. new ideas) as such may even originate from communication – think for instance of the quality-ameliorating virtuous cycle in customer-supplier feedback processes. Although these considerations apparently matter to the FP's intentions and layout, we will henceforth concentrate on spillovers since their relation to the FP was by far more investigated by economic authors.

Knowledge spillovers are the most frequently cited in favour of the FP. They pose a major disincentive to private R&D efforts: An organisation holding intellectual property may be subject to knowledge leaking out without receiving adequate compensation for it. Due to these positive externalities, the respective organisation is inclined to set its R&D investment lower than the social optimum. An increase of private R&D to socially desired levels may be induced by government intervention.

The very nature of spillovers in the European economy spurs the case for government action to be situated at the supra-national level: During the last decades, European integration and globalisation lowered barriers to international flows of goods, resources and knowledge; and the costs for communication and interaction decreased. This in turn implies growing importance of cross-border knowledge spillovers (Cornet 1999, p. 47). Most authors investigating the issue confirm international spillovers to exist and to be of considerable size. Inter alia, Coe/Helpman (1995, p. 872) show increases in foreign R&D capital stocks to affect domestic R&D and find the spillover effect to depend on bilateral trade. Moreover, they identify significant externalities even in the domestic market: According to Coe/Helpman (1995, p. 874) a 1% rise in domestic R&D capital stock would lead to a 1.23% increase in GDP for the G7 countries. In smaller OECD countries, this rate of return numbers only 0.85% due to higher openness to trade and international spillovers. The worldwide effect of a 1% increase (in R&D capital stock), however, would average 1.55%. This implies that at the domestic level research subsidies would only be effective in large countries, whereas a cooperative equilibrium in concerted international research policy would be mutually beneficiary. Regarding the integration among its member states, the international spillover effect can be reckoned to be particularly high in the European Union. This in turn calls for supra-national action at least at the continental level. Eaton/Gutierrez/Kortum (1998) support the case for a European R&D policy: They identify important cross-border spillovers and the small size of EU member states to constitute a disincentive to nationalised policy. Consequently, their study benchmarks EU data at the US and Japan, and estimate the marginal gain from an increase in research subsidies to be the highest in European countries (Ibid, p. 27).

Link (1981, p. 1112) and Funk (2002) assert that international spillovers are of particular relevance to basic research, which in turn provides a rationale for supra-national R&D policy to focus on pre-competitive research.

In theory, the issue of international spillovers may be addressed by market-based solutions as well: As illustrated in the surge of international RJVs, knowledge creators may form collaborative agreements by themselves in order to internalise knowledge spillovers. Moreover, a proper definition of intellectual property rights (IPR) may allow organisations to retain more benefits from knowledge production (Hall/Link/Scott 2001, p. 97). Although the latter argument points to IPR-related issues still to be resolved at the European stage, the intangibility of research outcomes hinders its contractibility, spillover internalisation by adequate pricing. The spontaneous set-up of international collaboration agreements is subject to massive information asymmetries. This relates to the difficulties in identifying appropriate partners, particularly in the case of cross-border and cross-culture cooperation. It may be argued that informational asymmetries decrease with the size of respective collaborating entities.

3.1.6 The Cohesion Motivation

While the prime motivation for the FP is to strengthen Europe's technological base, the general EU objective of "cohesion", (i.e. the support of less developed regions) is widely regarded as an important, and perhaps diluting, specificity of the FP. The two goals are frequently reckoned to conflict with each other, and numerous arguments have highlighted either its disturbing effect on project participation, or its long-term virtue in improving R&D skills in cohesion countries.

Sharp (1998) provides a thorough review of the cohesion effect's impact on FP policy, and, of particular interest to us, empiric insights into its influence on the regional distribution of FP3 funds.[39] With respect to policy, Sharp emphasises the positive effects from strengthening the R&D capabilities in cohesion countries. On the one hand she assigns the purpose of developing physical R&D infrastructure to EU structural funds (whose main objective is "cohesion"). On the other hand, Sharp (1998, pp. 581-582) finds FP collaboration to provide the opportunity to use these resources for learning (i.e. the transfer of tacit knowledge). With respect to this opportunity, Sharp (1998, p. 580) regards the diverting consequences of the cohesion effect as minor.

Sharp confirms the commonly asserted fact that cohesion countries (Greece, Spain, Ireland and Portugal) receive lower-than-average funds on a per capita or per GDP basis, while

[39] For confidentiality reasons, Sharp (1998) does not reveal the regional FP budget figures she uses.

small northern countries are allotted an over-proportional share. As an explanation, Sharp highlights the differing importance of the R&D over countries: While Sweden's total R&D expenditures amount to about 3.5% of GDP, the corresponding figure is around 0.5% for Greece and Portugal. This picture is further aggravated by the fact that public R&D expenditures account for more than two thirds of total in the latter two countries, while in Sweden it is about one quarter.[40] Higher education institutions in particular account for a large part of R&D in cohesion countries – accordingly, Sharp finds participants of those member states to be significantly more present in general science action lines as opposed to "technical" action lines like information technology or communication. She partly attributes this to the importance of large firms in the latter programmes, which vitiates against the participation of countries with few large leading edge firms in that sector.

When normalising FP fund distribution by the number of research staff, Sharp finds small states to fare better than large countries, with Greece, Ireland and Portugal receiving over-proportionally FP funds per researcher. In Greece and Portugal, FP3 funds added up to more than 10% of domestic civilian R&D expenditure. However, FP funds per researcher reveal a bleak picture in the case of Italy and, less so, in Spain.

The weak position of those two large, not centralised countries underscores the fact that cohesion is as much an issue within countries then among them. Hilpert (1992) identified 10 "innovative" islands among the EU-12 regions, and calculation based on our data set reveal that those regions provided for about half of EU-12 R&D expenditure while they accounted for roughly a third of EU-12 GDP.[41] Moreover Hilpert states that this concentration effect results in peripheral regions taking considerably less part in R&D collaboration schemes. However, European collaboration projects matter significantly more to peripheral regions than domestic schemes.

Sharp (1998, p. 583) finds that Hilpert's "innovative" regions dispose of 47% of EU-12 R&D personnel and received 47% of FP3 funds. Therefore she concludes FP funds to be distributed in a quite even fashion.

Based on literature, the effect of cohesion motivations on regional R&D collaboration is thus no entirely determinable. It certainly exists and induces more "cohesion" partners to participate than in a bottom-up equilibrium. But this may only partly offset concentration

[40] Figures based on own calculations. Data source: Eurostat (2003)

[41] Data Source: Eurostat (2003). Hilpert's (1992) "islands of innovation" roughly encompass the following NUTS-1 regions: DE1, DE2, DE7, DEA, FR1, FR7, ITC, NL3, UKH, UKI, UKJ. We related their 1996 R&D expenditure at the regional level against that of EU-12 at 1995 price parities.

effects between “innovative” core regions – thus it may move collaboration numbers closer to the patterns implied by geographic patterns of R&D resource distribution. Even if there may be a well-behaved distribution on the country level, the Commission only cares of cohesion *countries*. The marginal participations induced by the cohesion effect would otherwise not have been taken – and may therefore comprise the most advanced institutions in cohesion countries, which are generally located in core regions. Therefore the “cohesion effect” may even aggravate collaborative disparities within cohesion countries.

3.1.7 Conclusions from Institutional Characteristics

The vast majority of theoretic papers on knowledge generation and research collaboration strike the case for public intervention: Mainly arising from the quasi-public good character of knowledge, system failures and market failures beset the process of knowledge generation and related cooperation. Knowledge spillovers are the most frequently cited of these failures; asymmetric information and other symptoms obstruct the “bottom-up” contracting of such inefficiencies and thus result in sub-optimal R&D efforts. The public is requested to provide incentives in order to achieve the social optimum of R&D generation. In particular public-sponsored R&D collaboration is regarded as a means to increase “good” knowledge spillovers and compensate knowledge generators for their adverse individual effects. Since spillovers across intra-European borders have particularly risen, several authors justify the need for intervention on the supra-national level (although the origins of European R&D collaboration schemes did not ground on these considerations).

In line with the subsidiary principle, the FP evaluation criteria aim at correcting just for market failure on the European scale, while they nominally leave domestically feasible projects to the member states. The requested “additional value” attained through FP collaboration as compared to national schemes is particularly seen in large-scale, skill-demanding projects rather biased towards basic research. The latter stems from long-term strategy matters, from the need for public subsidies towards basic research, and from the fact that international spillovers occur particularly in basic research. Hence the FP is characterised as pre-competitive. Moreover, the evaluation criteria lay out demanding conditions for project applications, especially regarding the contribution to advancement in the field – those procedures are widely viewed as relatively bureaucratic.

The “pre-competitive” orientation of the FP and its bureaucratic processes demand highly developed scientific and administrative skills among prospective FP collaborators: This in turn could skew FP participation towards large corporate groups with interest in basic research and towards universities and research organisations. Smaller private entities, in contrast, are found to venture under-proportionally on FP projects, perhaps for lack of skills. The FP’s basic-research focus is reinforced through the competition by rivalling international collaboration schemes, most notably Eureka. The latter's competitive edge is in applied IT projects, and it seems to channel the demand for European near-to-market research – implying relatively less market orientation in the FP.

But the FP also competes with national collaboration schemes in the member states. In general, its importance there is fairly low. However, in the “Cohesion countries” the FP's significance is stronger, particularly since it is perceived to distribute funds “equally” among member states. The latter fact lies partly in the European Commission's practice to favour FP projects with “Cohesion country” participants – although the practice may widen internal disparities in the respective country.

The implications of the latter considerations will be accounted for in sections 5 and 6 and compared versus empiric results in section 7.

PART II – EXPLORATIVE STUDY

4 FP COLLABORATION: A SPATIAL INTERACTION PATTERN?

This paper aims at an explorative enquiry of FP collaboration data. For that purpose, we have the privilege to examine data newly compiled by Austrian Research Centers Seibersdorf (ARCS 2003). The database ranks among the largest on the topic and comprises virtually every project ever underdone in the FP structure. The range of dimensions offered by the data enables the investigation of many different aspects by a vast number of methods: Among these possibilities, this study focuses on the regional dimension of the FP4.

Concerning the organisation-specific, inter-relational and environmental factors affecting project participation and partner choice, regional data aggregation permits the investigation of the latter two. Sadly, this approach does hardly pay attention to factors particular to participating organisations. A profound analysis of individual participants is hard to perform, since organisational-level data is yet neither incorporated into the ARCS (2003) database nor otherwise freely available. Regional aggregation, in contrast, sets the focus on the linkage between spatial and local macroeconomic factors on the one side and FP cooperation on the other. Besides, geography is commonly regarded as important in FP collaboration: the majority of authors reviewing FP data investigate effects respective to the partner's countries. Apart from descriptive statistics, most of them include country dummy factors in their estimation procedures, which contribute significant explanatory value.[42]

In contrast to previous studies, this paper is (to our knowledge) the first to analyse not only country, but also inter-regional collaboration patterns[43] – and ranks among the few to analyse data at the aggregated level. (It has to be noted that we are able to perform valid tests on aggregated data only due to the large size of our sample). Moreover, we introduce the notion

---

[42] Inter alia: Fontana/Geuna/Matt (2003), Geuna (1998), Hernán/Marín/Siotis (2003), Marín/Siotis (2002), Navaretti et al. (2002).

[43] Sharp (1998) provides a comprehensive review on the regional allocation of FP funds and its effect on cohesion, but did neither focus on inter-regional collaboration nor on its "explanatory" factors.

of inter-regional FP collaboration as a pattern of spatial interaction. This renders the study's methods nearer to those of regional science and the economics of international trade than to the techniques followed by the research collaboration literature. In order to separate between spatial and environmental effects we rely on the most archetypal concept of spatial interaction models, the "gravity" model. By the means of this concept, we aim at explaining the matrix of inter-regional collaborative links in the FP4.

4.1 The Data Set

On the initiative of Dr. Roediger-Schluga, the underlying database was constructed at the ARCS's Department of Technology Policy. Its foundation rests on the juridical information freely available from CORDIS (2004), the European commissions main information site on EU innovation policy. During 2003, the whole of published data on the Framework Programmes 1 to 5 was collected and most of it categorised, standardised and augmented by additional information.

Basically, data is organised around FP projects. For each project, the database specifies the corresponding Framework Programme, the action line (up from the FP3), and the year of initiation. Moreover, a project contains data on its official members, including the name of the organisation, the country of its judicial seat, a responsible person and (in most cases) its address. Each participant in the project is categorised according to its organisational type and one of project partners is the designated prime contractor.

The data on the FP4 (1994-1998) constitutes the most recent sample offering high quality information on a completed FP. It exhibits cooperation at a larger scale than its predecessors, and provides sufficient standardisation and data quality. For that reason, this study will focus on data from the FP4.

4.1.1 The NUTS Nomenclature

For the purpose of this study, we achieved to describe cooperation at a lower level of geographic aggregation. In order to classify the region of a participating institution, we chose to use the unified structure of NUTS (Nomenclature des Unités Territoriales Statistiques). NUTS is a European standard to label the administrative subdivision of countries, where either NUTS-region is either a regional administrative entity or a sum of various entities. The so-called NUTS "levels" correspond to hierarchical administrative structure: For instance, the NUTS-0 label DE identifies Germany as a country. NUTS-1 regions correspond to German *Länder*, e.g. DE2 for Bavaria. The NUTS-2 level identifier DE21 is attached to one of its subdivisions, namely Upper Bavaria. And DE21H is the NUTS-3 level code for the district of

Munich, which positions it at the same rank in statistical hierarchy as a French Département.[44] For an overview over the NUTS-1 codes used in this study, please refer to Table A.1 in the appendix.

By drawing on postal addresses, we were able to attribute virtually all organisations with addresses to NUTS-1 level regions, and about 80% of them even to NUTS-3 regions.

In our estimation, however, we concentrate on collaboration between NUTS-1 regions: In this case, both the number of nodes and the scale of inter-regional cooperation are of sufficient size to perform the statistical methods used in this paper.

4.1.2 Excursion: Overview of Sample Aggregates

This study focuses on aggregate collaboration data organised along the regional dimension, but its scope omits to segregate for action lines or regarding specific organisational types. However, the participation per key action and the importance of various organisation types has been found to differ considerable between countries or regions (Sharp 1998). These differences in orientation may have a profound impact on the study of inter-regional collaboration. Since this diploma thesis is already considerably extensive and therefore voluminous, we will not account for these disparities henceforth. However, the issue is of importance and has to be kept in mind as holding explanatory power unaccounted for. Therefore we will provide a short overview of the dimensions omitted later on:

About 22,000 shared cost actions projects were carried out in the FP4 (European Commission 2001), counting more than 67,000 participants, i.e. about 3 participants per project. Our sample of the FP4 classes the projects and its adherents along the 13 key actions in the FP4's Activity 1 and the remaining three Activities.[45] Moreover, each participating organisation is classified either as an enterprise (36%), consultant (1%), a higher education institution (31%), governmental body (3%), private research organisation (26%), or as another non-commercial institution (0.4%).[46]

---

[44] For further details, refer to European Union (2003).

[45] The two action lines "controlled nuclear fusion" and "nuclear fission safety" are not included in the database, since they adhere to the Euratom Framework Programme rather than to the "core" FP (compare Table 2, p. 57).

[46] Shares of the respective organisational type in the sample are in parentheses. Figures do not add up to 100% because of observations not already attributed to a specific organisational type.

Of the 67,682 participants in the sample, 92.3% have been attributed to a NUTS-1 region, and 85.7% of all participants were identified to be located in one of the 72 NUTS-1 regions[47] of the EU-15. Those 85.7% were taken to construct the matrix of inter-regional collaborative links, which constitutes the data set we aim to explore and explain in this study.

Information by CORDIS rendered it possible to divide projects into action lines. With 21% of the total, the Advanced Materials key action numbers by far the most participants, followed by information technology and by Activity 2 and Activity 4, each accounting for about 10% of participants. It has been widely noticed that the key actions are subject to wide disparities in the participation of private firms: While in socio-economic research and Activity 4 firms' share of participants is less than 6%, it numbers about 55% for standards & measurements and transport, and more than 65% for non-nuclear energy, communication, and information technology. In particular the information and communications technologies (ICT) key actions are regarded as more oriented towards applied research (Luukkonen 2002), whereas those with less frequent industrial participation are rather interpreted as focused on basic research. It is hypothesised that the poorer EU regions' universities have a comparatively strong standing, whilst their private firms are less interesting preferred partners in European applied research. However, the data set we are analysing does not support this hypothesis. For instance the four "cohesion" countries" Spain, Greece, Ireland and Portugal together exhibit 15% of their participations in ICT projects versus 13% on average.

ICT and Collaboration Intensity

Our data set hints at a weak positive relation between FP4 project participations per capita and the importance of ICT key actions, which both are figures of considerable variance. But as Sharp (1998) points out, the "potential" for regional total collaboration may rather be the number of researchers rather than the entire population. Figure 2 displays European NUTS-1 regions[48] according to the share of ICT participations (ICT intensity) and total collaboration per employed researcher (collaboration intensity). We leave the analysis of explanatory factors for total collaboration to sections 5 and 6, but will use data on FP4 key actions to visualise some interesting disparities in regional FP4 numbers in Figure 2: "Rich" regions appear at the centre of the graph, with Northern Italy, Paris and several English regions (most notably London and the "Oxbridge" area") ranking ahead of their national counterparts in ICT cooperation. Besides, these regions surpass their domestic peers considerably when

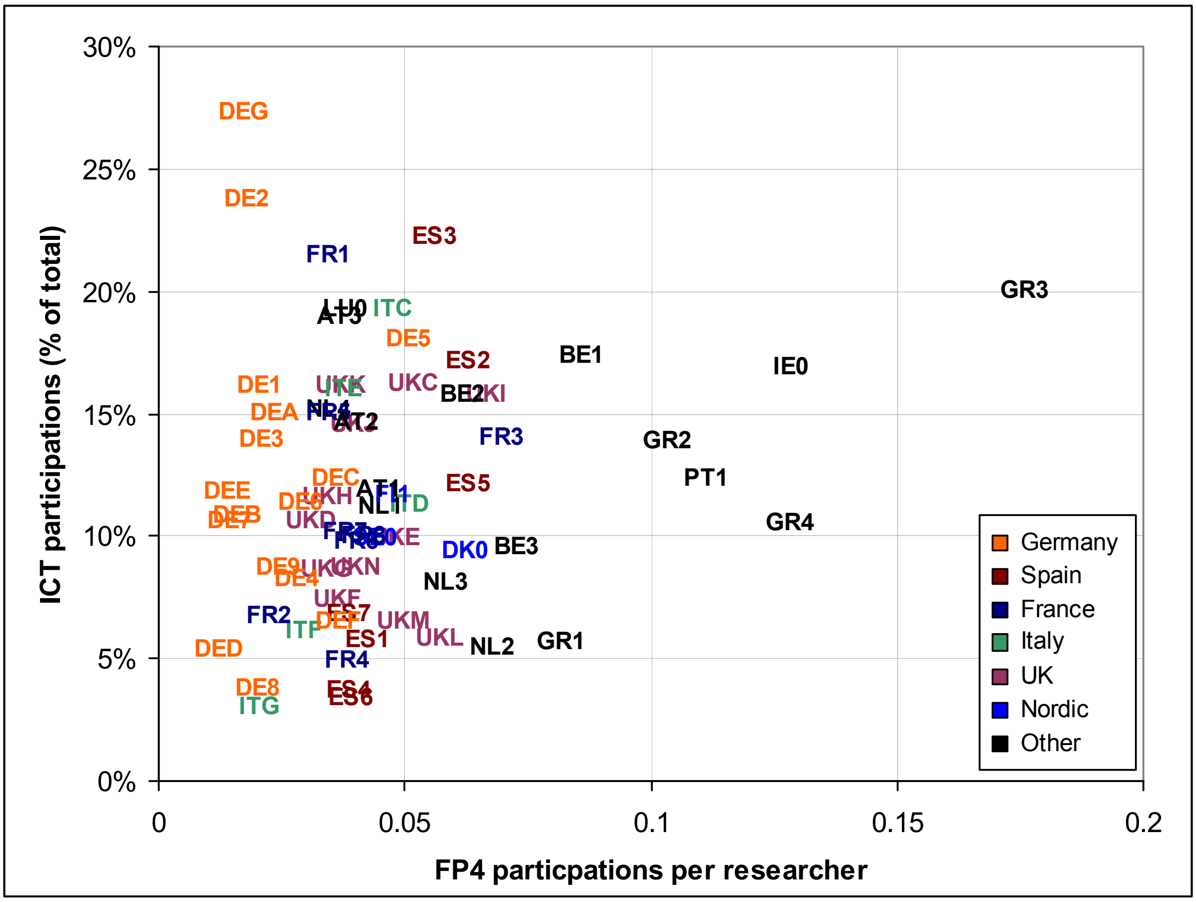
[47] The number of 72 regions corresponds to the state in 2004, previously lowered through the recent revisions of Italian and British NUTS-1 regions.

[48] See Table A.1 in the appendix for a key of NUTS-1 codes.

looking at project participations per capita. The German regions occupy the points farthest to the left, but figure above average in ICT cooperation, which is even more pronounced for the industrial heavyweights DE1, DE2 and DEA. Interestingly, the sparse area to the right is held by Greece (particularly its capital GR3), Ireland and Portugal. All of them are “cohesion” regions, and all are over-proportionally involved in ICT collaboration. The latter is even more important for the industrialised Spanish regions, namely ES2 (the North-East), ES3 (Madrid) and ES5 (Catalonia). In contrast, several “cohesion” regions hold the lower left end, mainly Eastern German *Länder* and the rural, poorer regions of Spain, Italy and also France. Interestingly, the share of ICT in total participations is considerably lower than average for the Scandinavian countries, the Netherlands and the entire UK bar the Southeast. Besides, it appears that researchers from small countries (in black colour) seem more committed to FP participation than their colleagues in larger EU member states (represented by coloured points). In particular Belgium seems to be more intensely involved into FP cooperation than regions of comparable GDP per capita

Particularly when regarding “cohesion” countries, the question arises why some of them seem to excel in FP collaboration, while others appear as under-performers even when related to their smaller resources? The sections to follow try to answer this question inter alia.

Figure 2: FP4 Collaboration per researcher versus participation in information and communication key actions



Participation by Organisational Type

Table A.3 in the appendix presents the importance of organisational types in FP4 collaboration. By construction, the regional proportion of ICT collaboration is related to the share of private firm participants. As may be expected based on Figure 2, the highest firm shares are to be found in Germany, but also France and Spain. The Nordic countries and the Netherlands, in contrast, feature considerably less firm participation compared to the average and even to “cohesion” countries. In each country the dominant part of non-firm participants divides into research organisations and higher education institutions, with a trade-off in the shares of the latter two organisational types. The proportion of education institutions in total participants is particularly low in France, while it is substantially above average in Ireland, Belgium, Sweden and Spain. Conversely, the latter four countries exhibit low shares of research organisations, while these organisations matter most in France, Italy, Finland, Denmark. Note the difference between Sweden on the hand and its Nordic neighbours on the other. Along the triad of firms, universities and research organisations, governmental bodies are less but still important, while consultancies or private non-profit institutions play a minor role.

Sector Collaboration Intensity

Furthermore, the intensity of FP4 collaboration with respect to organisational types appears to differ strongly when put in relation to available resources: In the EU-15, business sector researchers account for nearly 60% of total research staff, but industrial participation in the FP4 is only 36% of the total. We performed a rough calculation, dividing industrial, educational and research organisation participants (plus governmental bodies) by the Eurostat (2003) research staff numbers for the business, education and public non-educational sector. In total, 25 out of 1,000 business researchers participated in the FP4, while the corresponding figure is 46 for research organisations, and even 75 for educational institutions. Thus the more a region is inclined towards public research the more it apparently collaborates.

For most of the regions considered, the distribution among sectors follows broadly the relations presented. There are exceptions, though: Most notably, “educational” participations number more than 200 per 1,000 educational researchers in Ireland, and still more than 100 in UKH (“Oxbridge”), Greece and BE3 (Walloon Region). Industrial participations reach about 300 per 1,000 business researchers in Greece and Portugal. Research organisations, in contrast, boast several hundreds per 1,000 in AT2, Belgium, FR3, FR4, Ireland, NL4 and UKC[49] – and are considerably above average throughout French, Northern Italian, Dutch and English regions.

The Role of Prime Contractors

The core-periphery presumption re-emerges to some extent when looking at the distribution of prime collaborators. As laid out on the following pages, the prime contractor can be regarded as the “centre” of an FP project, thus benefiting the most from potential positive effects due to knowledge spillovers and absorption. When relating the number of prime collaborators to a region’s project participants, the four “cohesion countries” exhibit a clearly lower share than the bulk of other countries.[50] Britain, and to a lesser extent the Benelux countries and the Paris-Lyon-Marseille corridor of France exhibit the highest prime contractor shares. But the proportion of prime contractors among total researchers per regions shows once again the Greek and Irish regions topping the list, together with London, Belgium and the Netherlands. German regions share the lowest ranks only with Southern Italy and

---

[49] Compare Table A.1 in the appendix for the corresponding region names.

[50] Only the “newcomers” Sweden and Finland exhibit prime collaborator shares comparable to the cohesion countries Spain, Greece, Ireland and Portugal.

Southern Spain. The remaining regions exhibit average prime contractors per researcher, (most notably Paris figuring among them).

Concluding Remarks

Concluding, the data presented confirms the familiar picture of the UK, Belgium, the Netherlands and France as being the “centres” of FP4 interaction. The “cohesion” countries, Greece, Ireland and Portugal appear intensely involved into FP4 cooperation, with average shares of firm participants and “applied” projects. German collaboration seems subject to particularities: Especially German public organisations seem under-represented in FP4 collaboration. This translates into overall German underperformance, given that non-firm institutions account for nearly two thirds of FP4 participants. Furthermore, we identify fierce intra-country disparities adversely affecting the weaker German, Spanish and Italian (and French) regions. On top of modelling FP interaction, similar disparities will be further enquired in the sections to follow. This rough overview already hints to many facts and questions of interest, but their further investigation would unfortunately exceed the scope of a diploma thesis.

4.1.3 Assumptions Concerning Network Structure

CORDIS provides us with no information on the internal structure of the projects. Several studies based on the INNOCULT survey (e.g. Pohoryles 2002, p.333) try to determine how many of the observed projects are organised by “hierarchic”, “individual” or “communitarian” structures. For that purpose, it is helpful to depict the internal project partners as a network. In this case, the partners constitute the nodes while the lines between them represent the degree of interaction. In various fields of research, it is common to analyse directional connecting lines – e.g. in trade economics, the graph between two countries separates into two vectors illustrating import and export flows. In the case of complex research collaboration, however, it is hard to determine a (net) direction of informational flows. Knowledge generation in this case stems from information exchange, not delivery. Therefore it is reasonable to assume that in our case cooperation among partners is not directional. Furthermore, it is nearly infeasible to determine the intensity of interaction (the weight of the connecting line) among project partners for a large data sample. For this reason, we assume all connections to be of equal intensity.

Although we are able to identify the prime contractor among project partners, we cannot determine how the internal project structure is organised. We face the choice between two extremes: Either we assume the project to follow a radial structure, where all parties to the project only cooperate via the prime contractor. Or we interpret the project as “communitarian” which implies a fully connected network. We choose the latter alternative, since Pohoryles (2002, p.334) finds the dominant part of projects to follow "communitarian" or "individualist" patterns: Although those styles differ in emphasis on interaction, both attribute rather equal importance to nodes.

On the aggregated level, our assumptions imply that a project of n adherents is counted as a network with nodes of equal properties (apart from organisation type and region) and with $\frac{1}{2}n\,(n-1)$ undirected connecting graphs.

Regional aggregation omits the project dimension and focuses on the number of intra-project links between different regions. Table 3, for instance, represents a summary of cooperation links among EU-15 member states.

Table 3: Intra-project cooperation between EU-15 project partners in our data set[51]

	AT	BE	DE	DK	ES	FI	FR	GR	IE	IT	LU	NL	PT	SE	UK
AT	712	461	2,202	364	621	358	1,102	351	210	1,038	31	705	228	478	1,306
BE	461	1,567	3,350	630	1,290	591	3,396	762	436	1,930	96	1,963	622	847	3,322
DE	2,202	3,350	13,898	2,271	4,450	2,240	10,493	2,394	1,115	7,136	141	5,633	1,692	3,552	11,135
DK	364	630	2,271	683	793	587	1,619	539	319	1,142	37	1,327	379	1,055	2,540
ES	621	1,290	4,450	793	4,159	742	4,742	1,524	610	3,851	46	1,873	1,616	1,335	4,870
FI	358	591	2,240	587	742	751	1,424	593	326	1,118	32	1,043	357	991	2,024
FR	1,102	3,396	10,493	1,619	4,742	1,424	10,817	2,241	1,111	7,213	116	4,013	1,705	2,524	10,563
GR	351	762	2,394	539	1,524	593	2,241	1,536	362	2,360	39	1,129	672	648	2,857
IE	210	436	1,115	319	610	326	1,111	362	364	805	20	609	329	430	2,094
IT	1,038	1,930	7,136	1,142	3,851	1,118	7,213	2,360	805	5,890	45	2,687	1,340	1,759	7,121
LU	31	96	141	37	46	32	116	39	20	45	30	48	26	41	83
NL	705	1,963	5,633	1,327	1,873	1,043	4,013	1,129	609	2,687	48	3,827	770	1,673	5,868
PT	228	622	1,692	379	1,616	357	1,705	672	329	1,340	26	770	722	510	2,146
SE	478	847	3,552	1,055	1,335	991	2,524	648	430	1,759	41	1,673	510	1,271	3,692
UK	1,306	3,322	11,135	2,540	4,870	2,024	10,563	2,857	2,094	7,121	83	5,868	2,146	3,692	15,012

If we regard the matrix in Table 3 as depicting a network, its columns, respectively its rows represent n=15 nodes, while the values illustrate the intensity of cooperation between every possible pair of nodes. The assumptions of no directionality imply the matrix to be symmetric along its main diagonal: There are $\frac{1}{2}n\,(n-1) = 120$ possible combinations of different nodes i, j plus 15 intra-regional interaction intensities, in sum $\frac{1}{2}\,n\,(n+1) = 135$ values illustrating interaction.

[51] Note: The values of FI exclude the contribution of Åland Islands, those of FR exclude the French overseas Départements and those of PT the Madeira and Açores Islands. The data set includes only observations attributed to a NUTS-1 region – therefore numbers do not exactly match data given by the European Commission (1998).

The displayed geographic cooperation structure provides the basis for the following sections of this study. The statistical methods employed henceforth aim at understanding and explaining the such-defined matrix.

4.2 The "Gravity" Model

Interpreting the cooperation matrix as a spatial interaction structure reminds of similar patterns in regional science (e.g. commuting flows) or trade economics (e.g. import/export flows). One of the widely known methods to analyse spatial interaction is known as "gravity" model, after the Newtonian concept of gravity.[52] By analogy, the potential of interaction T_{ij} between nodes i and j is dependent on properties particular to each node M_i and M_j (their masses, in Newtonian mechanics) and on their relative positioning (Distance) D_{ij}.

(1) $$T_{ij} = \mathrm{f}(\underset{(+)}{\theta}, \underset{(+)}{M_i}, \underset{(+)}{M_j}, \underset{(-)}{D_{ij}}, \varepsilon_{ij})$$

The relationship displayed in (1) closely follows the Newtonian analogy. T_{ij} is dependent on a function positively affected by masses M_i, M_j and inversely related to distance D_{ij}. Furthermore a uniform constant θ provides for a scaling of the function (Newton's gravitational constant). In addition to the deterministic Newtonian concept, we introduce an error term, since social science data is stochastic by nature.

We develop (1) further to the structure in (2), which resembles even more the familiar formula from physics.

(2) $$T_{ij} = \theta M_i^{\alpha} M_j^{\alpha} D_{ij}^{-\beta} \mathrm{e}^{\varepsilon_{ij}}$$

Although the multiplicative structure of (2) follows rather straightforward out of analogy,[53] particular attention has to be paid to the structure of the error term ε_{ij}. The exponential

---

[52] Note: The economic disciplines relying on this type of model have extended and generalised the concept. They frequently avoid the term "gravity" and rather describe it as a model of "spatial interaction". Comprehensible, since the general "spatial interaction" models may as well be based on thermodynamics.

[53] It has to be noted, however, that most gravity approaches in social sciences attribute differing origination and attraction functions to the masses M_i and M_j, expressed in differing exponentials α_1 and α_2. Since they mostly study directional interactions, i.e. $T_{ij} \neq T_{ji}$, they may assume different mass effects on outflows versus inflows. Our data set, in contrast, consists of symmetric interaction and

assumes ε_{ij} to be independent of other variables in the multiplicative structure in (2), as well as in the log-transformation of this structure (3).

(3) $\ln(T_{ij}) = \ln(\theta) + \alpha \ln(M_i) + \alpha \ln(M_j) - \beta \ln(D_{ij}) + \varepsilon_{ij}$

The additive structure of (3) displays the same relationship as (2), but its linearity due to the natural logarithm facilitates the application of standard methods. Moreover, the error term ε_{ij} remains a pure and single summand. This in turn helps in the analysis of statistical properties.

Moreover log-transformation enables compressing the total of the interaction matrix into a simple formula (4).

(4)
$$\begin{aligned} &\ell\mathbf{T} = \ln(\theta)\vec{\mathbf{1}} + \alpha\,\ell(\mathbf{m}\times\mathbf{m}') - \beta\,\ell\mathbf{D} + \boldsymbol{\varepsilon} \\ &\text{where} \qquad \ell\mathbf{T} = \{\ln T_{ij}\},\ \vec{\mathbf{1}} = \{1\},\ \ell(\mathbf{m}\times\mathbf{m}') = \{\ln(M_i M_j)\}, \ell\mathbf{D} = \{\ln(D_{ij})\}, \boldsymbol{\varepsilon} = \{\varepsilon_{ij}\} \end{aligned}$$

The $n \times n$ logarithmic transaction matrix $\ell\mathbf{T}$ depends on the scalar θ, on the outer product of $n \times 1$ mass vector $\mathbf{m}$ (times a coefficient α), on the $n \times n$ logarithmic distance matrix $\ell\mathbf{D}$ (times a coefficient β), and on the $n \times n$ error matrix ε. [54]

In contrast to most economic applications of the gravity concept, the non-directionality of our data set renders the corresponding matrices symmetric along their main diagonal ($T_{ij} = T_{ji}$). Moreover, we find no reason to assume the distance between node i and j D_{ij} to differ from the distance between j and i D_{ji} – which implies that $\varepsilon_{ij} = \varepsilon_{ji}$. This property leads to some convenient implications, notably to the fact that we only need to compare the main diagonals and the lower triangular matrices of $\ell\mathbf{T}$, $\ell(\mathbf{m} \times \mathbf{m}^T)$, $\ell\mathbf{D}$ and ε.

In order to apply standard least squares (LS) estimation techniques to spatial interaction data, it is important that interaction is sufficiently frequent to attain a relatively continuous distribution of observations. Many spatial interaction data sets similar to the ours exhibit low frequency and thus follow a rather discrete pattern which calls for the use of probit models, GMM estimation or similarly elaborated techniques. Based upon performance in Monte Carlo simulation, Sen/Smith (1995, pp. 519-525) assess the quality of OLS to be sufficient if

---

distance matrices ($T_{ij} = T_{ji}$, $D_{ij} = D_{ji}$); therefore the mass exponentials are set equal to each other. In the present model structure, this can only hold if both mass exponentials are equal.

[54] The ℓ operator adjacent to a matrix $\mathbf{X}$ denotes a matrix whose elements consist of the logarithms of the elements of $\mathbf{X}$ in the same position. This definition is necessary since the “logarithm” of a matrix as such is not defined.

bilateral interaction values T_{ij} are large and n is “not too small”. Although they do net set clear boundaries, they reckon a T_{ij} larger than 3 leading to unbiased results with estimating log-transformed implied distance (Ibid., p. 481).[55]

The properties of our collaboration matrix seem to fulfil these requirements: It has to be noted, that the aim of Sen/Smith (1995) is to obtain accurate estimates for θ. Our purpose, in contrast, is to evaluate the relevance of several prospective explanatory factors and the sign of their impact rather than their precise magnitude. Out of the guidelines we were able to consolidate from Sen/Smith (1995), we deem the fallacies to OLS estimation relatively minor in our case. Moreover, only 65 out of our n (n+1) /2 = 2346 observations are smaller than 3.

Apart from the sufficient size, our data matrix and explanatory variables have as well to conform to several other conditions, notably that the variables under consideration are normally distributed, while the error terms should have a mean of zero, and ought to be independently identically (normally) distributed. The latter condition in particular requires error terms to be uncorrelated, thus $Cov(e_{ij}, e_{lm})=0$, $i \neq l$, $j \neq m$. This problem will be examined in further detail during section 6.1.

5 Exploration of Regional Collaboration Data

The remaining part of this diploma thesis is dedicated to the empirical endeavour of collaborative links among 68 regions in the Fourth Framework Programme. The agglomerate data can be consolidated into a symmetric matrix, which we assume to display a network arising from spatial interaction. In this context, the next pages present a deterministic transformation of the collaboration matrix in order to analyse several properties against the backdrop of the mass-distance dichotomy introduced before. The ultimate goal is to assess the appropriateness of a gravity model in explaining the matrix’s interaction pattern. Implicitly drawing on the theoretic considerations laid out in sections 2 and 3, we introduce prospective explanatory factors for the gravity model structure. However, since the character of this diploma thesis is rather exploratory, we later on try to handle data even-handedly – i.e. without placing too much restriction on variable choice. Instead, we focus on the design of the variable selection procedure (section 6.1.4), and only afterwards we assess the model’s intuitive value with respect to suggestions by literature (section 7).

[55] In Sen/Smith (1995, p. 458) the authors add ½ rather than 1 prior to log-transformation.

5.1 The Dependent Variable: Inter-Regional Research Collaboration

Out of our regionalised data set, we were able to compile a regional cooperation matrix describing cooperation between European NUTS-1 and NUTS-2 regions of all European countries taking part in the FP4. Apart from the EU-15, the Central and Eastern European Countries are well represented. Nevertheless, during 1994-1998, they still exhibited much lower cooperation numbers than their western counterparts – mainly due to the fact that during this period east-west integration was still in its initial phase. Furthermore, and despite of the already vast amount of Eurostat data on candidate countries, prospective explanatory regional data still is not up to the EU-15 in terms of range and quality of explanatory variables. Adding Switzerland and Norway to the EU-15 data seemed a better option, since both parties already played an important role in EU cooperation during the nineties. However, since most of our explanatory data stems from Eurostat, we lack standardised information on some of the explanatory variables available for EU-15 regions. Therefore we confined ourselves to the European Union within the borders of 1995. For these 15 countries, explanatory data on the NUTS-2 level is easily accessible, albeit it bears some data gaps. Data on NUTS-1 regions is slightly more complete. Even more important, a considerable number of NUTS-2 regions provide too infrequent cooperation numbers to perform valid estimation techniques. Therefore we chose the regional dimension to be represented by the 72 NUTS-1 regions in the EU-15. (Please see Table A.1 in the appendix for a key of NUTS-1 codes and region names).

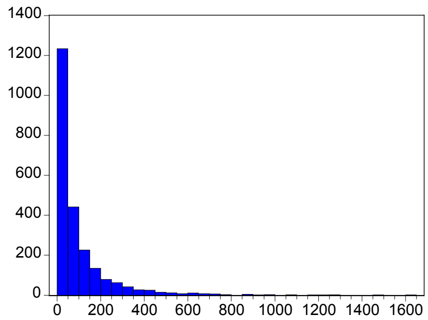
The average cooperation between each pair of those exhibits a rather large scale with a mean of 99 inter-regional collaboration links and a median of 49. However, we had to exclude four autonomous or semi-autonomous regions: The Finnish Åland Islands (NUTS-1 code FI2) the French overseas Départements-d'outre-Mer (FR9), and the Portuguese islands of Açores (PT2) and Madeira (PT3) exhibit by far lower cooperation numbers than all of the other NUTS-1 regions in the sample. The exclusion of the latter four regions leaves us with a cooperation matrix of 68 nodes, each representing a region. Collaboration linkages in this matrix sum up to 441,329. The node with the most collaborative linkages is FR1 (Île-de-France / Paris Region), with a total of 27,901 links to the 68 nodes, while the minimum sum of collaborations is found in ES7 (Spain's Canarias) with 586 linkages.

The matrix contains $\frac{1}{2}$ n (n+1) = 2346 possible combinations of nodes. The largest single value (1,647) is that of FR1 cooperation with itself, the largest inter-regional value (1,490) that of FR1 with ITC.[56] Only 1.1% of matrix components are limited to zero, while 85% display at least a value of 10.

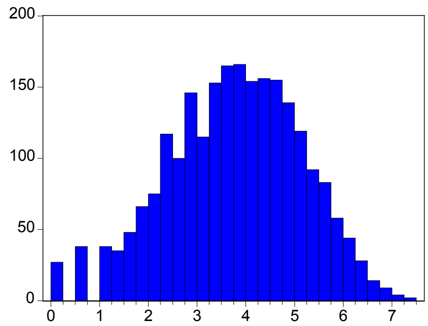
The left part of Table 4 displays the histogram and a selection of descriptive statistics for the 2,346 cooperation values between all 68 nodes. They seem to follow a rather exponential distribution, and the Jarque-Bera test on normality clearly rejects the hypothesis of normal distribution. The right-hand part of Table 4 depicts the natural logarithm of the left values plus one (the "log-transformed" cooperation data).[57]

Table 4: Histograms and descriptive statistics of cooperation between NUTS-1 regions in the fourth Framework Programme – standard and logarithmic

<table>
<tr><td>CoopFP4:
Cooperation between 68 NUTS-1 Regions in the FP4 (main diagonal and lower triangular values)</td><td>L_CoopFP4:
Logarithm of the cooperation between 68 NUTS-1 Regions in the FP4 plus one (main diagonal and lower triangular values)</td></tr>
<tr><td>Histogram CoopFP4



</td><td>Histogram L_CoopFP4



</td></tr>
<tr><td>Descriptive Statistics
Observations 2346

Mean 97.50469
Median 45.00000
Maximum 1646.000
Minimum 0.000000
Std. Dev. 147.4814
Skewness 3.757174
Kurtosis 23.87943

Jarque-Bera 48133.68
Probability 0.000000</td><td>Descriptive Statistics
Observations 2346

Mean 3.763354
Median 3.828641
Maximum 7.406711
Minimum 0.000000
Std. Dev. 1.381296
Skewness -0.254111
Kurtosis 2.777350

Jarque-Bera 30.09357
Probability 0.000000</td></tr>
</table>

The data in the logs bear rather more resemblance to the Gaussian distribution and its Jarque-Bera-statistic is fairly low. Nevertheless, Jarque-Bera still rejects the hypothesis of

[56] The Italian Nord-Ovest, comprising Lombardia, Piemonte, Liguria and Vale D'Aosta. Both FR1 and ITC rank among Europe's top five regions in terms of population and GDP.

[57] Because the logarithm cannot be applied to values of 0, we add 1 to the data to achieve smooth logarithmic transformation. In order to ensure compatibility, we perform the same procedure on all logarithmic transformations mentioned in this study.

normal distribution. By the analysis of Q-Q plots (not displayed), we conclude that this is due to the "fat" left tail of the histogram - i.e. the number of very low values in the matrix is slightly too high.[58] Regarding the "near-normality" of the sample, we judge the data properties as sufficient with respect to statistical requirements.

5.1.1 Principal Components Decomposition

The cooperation matrix' symmetric properties and its full rank allow for a reduction of its dimensions: In order to facilitate interpretation, we decompose the matrix T into eigenvalues and eigenvectors. The procedure consists of creating a transformation matrix P that follows the conditions described in (5).

(5) $$\mathbf{P\,T\,P'} = \mathbf{\Lambda},\quad \mathbf{P} = \begin{pmatrix}\mathbf{p_1} & \dots & \mathbf{p_n}\end{pmatrix},\quad \mathbf{\Lambda} = \begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \lambda_n\end{pmatrix}$$

(6) $$\mathbf{P}^{-1} = \mathbf{P}',\quad |\mathbf{p_i}| = 1,\quad \mathbf{p_i'}\mathbf{p_j} = 0 \qquad \forall\, i, j$$

The n x n transformation matrix **P** consists of n 1 x n eigenvectors $\mathbf{p}_i$, and the diagonal elements of Λ are the eigenvalues λ_i corresponding to the eigenvectors $\mathbf{p}_i$. The inverse of **P** is required to match its transposed equivalent **P'**. Moreover its components, the eigenvectors, are required to have norm 1 and to be exactly orthogonal to each other (6). Thus each eigenvector represents exactly one of **T**'s n dimensions. According to (5) and (6), the transformation matrix can be determined by solving for (7).

(7) $$\mathbf{P\,T\,P'} = \mathbf{\Lambda} \;\Rightarrow\; \mathbf{PT} = \mathbf{\Lambda P} = \mathbf{P'\Lambda} \;\Rightarrow\; \mathrm{Det}(\mathbf{T} - \mathbf{\Lambda}) = 0$$

Following (5), one may transform into the representation displayed in (8).

(8) $$\mathbf{P\,T\,P'} = \mathbf{\Lambda} \;\Rightarrow\; \mathbf{T} = \mathbf{P'\Lambda P} \;\Rightarrow\; \mathbf{T} = \lambda_1\mathbf{p_1'}\mathbf{p_1} + \dots + \lambda_i\mathbf{p_i'}\mathbf{p_i} + \dots + \lambda_n\mathbf{p_n'}\mathbf{p_n}$$

Eigenvalue decomposition has the convenient property that the eigenvector $\mathbf{p_i}$ corresponding to the largest eigenvalue λ_1 (the first principal component) captures the maximum variance of T possible to explain by one dimension. The eigenvector adjunct to next largest eigenvalue captures the maximum amount of the remaining variance not explained by the first principal component since it is exactly orthogonal to the former eigenvector.

[58] Moreover, a 1-digit difference at the very low end of the range transforms into by far higher differences of their logs than in the rest of the sample. This leads to digital behaviour of values at the lower end, which explains the gaps at the histograms' very left tail.

In this manner, we are able to construct n independent n x 1 vectors describing the total of matrix T, while a combination of a few of these vectors already represents a sufficient approximation of T. Table 5 depicts the 12 largest eigenvalues (in absolute size) of the cooperation matrix T and of its log-transformation.

Table 5: Largest Eigenvalues of Cooperation Matrix[59]

Rank		Cooperation Matrix T				Log-transformed T		
		Eigenvalue	Explanatory Power . R^2 component matrix	Cumulative Explanatory Power		Eigenvalue	Explanatory Power R^2 component matrix	Cumulative Explanatory Power
1		11,395.66	93.0%	93.0%		269.75	88.7%	88.7%
2		1,469.26	2.3%	95.2%		13.80	2.2%	90.9%
3		1,119.23	1.3%	96.6%		-11.62	1.5%	92.4%
4		-743.45	0.6%	97.2%		11.28	1.5%	93.8%
5		729.09	0.6%	97.7%		8.62	0.8%	94.7%
6		639.64	0.4%	98.2%		6.94	0.5%	95.2%
7		517.44	0.3%	98.4%		6.77	0.5%	95.8%
8		458.25	0.2%	98.7%		6.20	0.4%	96.2%
9		382.08	0.2%	98.8%		5.64	0.4%	96.6%
10		-375.27	0.1%	99.0%		5.25	0.3%	96.9%
11		351.37	0.1%	99.1%		4.87	0.3%	97.1%
12		-320.47	0.1%	99.2%		-4.68	0.2%	97.4%

The largest eigenvalue of the log-transformed **T** already displays an R^2 of 93% to matrix **T** (88.7% in the non-transformed case). This implies that the 1-dimensional first eigenvector describes a dominant part of the n-dimensional cooperation matrix.

Figure 3 displays eigenvector $\mathbf{p_1}$ and the total number of collaborative links per region (the row sums respectively column sums of the cooperation matrix). It is striking how closely both vectors fit together. Since row/column sums of a spatial interaction matrix are often regarded as an indicator of gravitational masses, the first principal component thus is regarded as a "masses" proxy, adjusted for a certain centrality-periphery component.

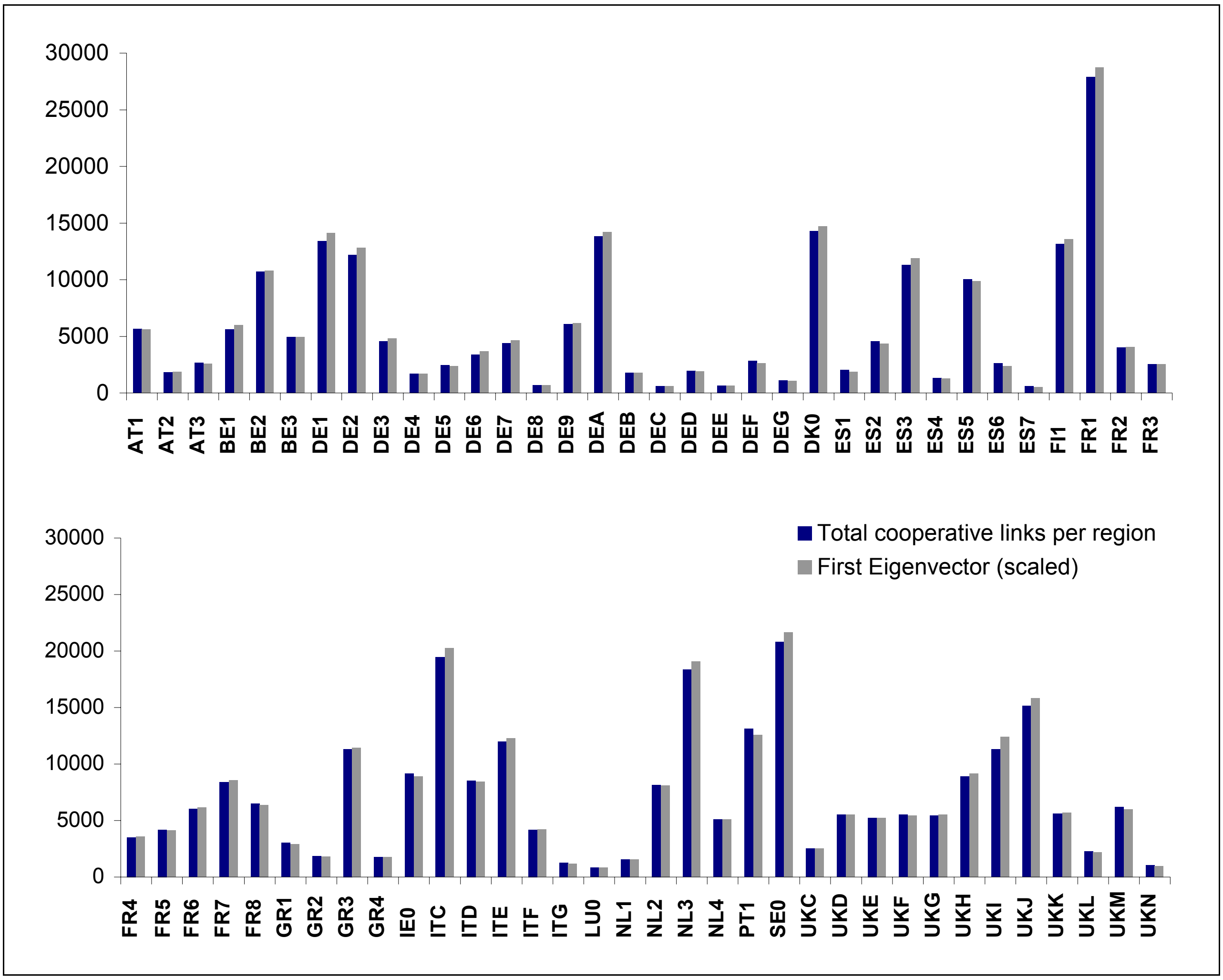
The eigenvectors corresponding to the second and third largest eigenvalues are harder to interpret: They nearly cancel each other out for about half of the regions, but adjust for several regions in order to describe cluster effects: DK0 (Denmark), for instance, is by far stronger involved in cooperation with the Nordic countries and the Netherlands than the product of the respective row sums would imply – collaboration with the Romanic countries, in contrast, is considerably less pronounced than the first principal component would suggest. Thus we conclude that the second and third principal component already catch

---

[59] Eigenvalues are ranked by absolute size. "Explanatory Power – R^2 component matrix" is the R^2 between the collaboration matrix and a matrix constructed out of the outer product of eigenvalue's corresponding eigenvector. "Cumulative Explanatory Power" is the R^2 between the collaboration matrix and the sum of the x most important eigenvalues and their vectors' outer products as in (8).

relational aspects of FP cooperation. But along with a distance matrix, the outer product of a mass vector (with itself) constitutes as well a relational factor.

Figure 3: Scaled eigenvector (of the first principal component) and total number of collaborative links per NUTS-1 region



The eigenvector decomposition allows for a division of matrix effects into uncorrelated vectors, which in turn eases the usability analysis of potential explanatory variables – please see p. 115 on this topic.

5.1.2 Relative Positioning – Implied Distance Matrix

The regional cluster effects just illustrated for Denmark may be observed with many regions: With some regions, a NUTS-1 entity under investigation may collaborate more intensely than it “should” according to gravitational mass effects, while there may less than “normal” cooperation with other regions. If **T** is interpreted as a spatial interaction pattern, those relational factors not attributable to n x 1-dimensional masses are due to differences in node-to-node-distance. Sen/Smith (1995, p. 478) provide a simple procedure to adjust for masses and transform **T** into a matrix of implied Distances **D***.

Ignoring error terms, the authors base their transformation on the simple gravitational model in (9), with all entering matrices implicitly being symmetric along their main diagonal.

(9) $T_{ij} = \dfrac{\theta M_i^\alpha M_j^\alpha}{\mathrm{f}(D_{ij})}, \qquad D_{ij} = D_{ji}\ ,\ T_{ij} = T_{ji}\ ,\ D_{ii} = 1\ \ \forall\, i,j$

Furthermore, the "internal" distance determining intra-regional collaborative links is normalised to 1. This assumption may be justifiable regarding that Eurostat classification requires NUTS-1 region boundaries to encompass about an equal number of inhabitants. The representation in (10) follows suit out of (9):

(10) $\dfrac{T_{ij}\ T_{ji}}{T_{ii}\ T_{jj}} = \dfrac{\mathrm{f}(D_{ii})\,\mathrm{f}(D_{jj})}{\mathrm{f}(D_{ij})\,\mathrm{f}(D_{ji})} \quad \Rightarrow \quad \dfrac{\sqrt{T_{ii}T_{jj}}}{T_{ij}} = \mathrm{f}(D_{ij}) \qquad \mathbf{D}^* = \{\mathrm{f}(D_{ij})\}$

The implied distance matrix **D*** consisting of components constructed from (10) thus represents a mass-adjusted indicator for relative collaboration impediments.[60] This facilitates the analysis of the relationship between node i and all other nodes. However, the **D*** of the transaction matrix boasts a rank of n=68, complicating the illustration of multi-node patterns. Therefore we apply once again eigenvector decomposition in order to reduce dimensions to 2 with the intention of displaying the relationship graphically.

Basically, the procedure draws on the (implied) distance matrix in order to generate a matrix of coordinates of similar dimensions. Coordinate vectors are based on eigenvectors, and are ranked by the largest eigenvalues (in absolute size). This allows for depicting the maximum of distance matrix variance in reduced dimensions. Mardia/Kent/Bibby (1995, p. 400) describe a simple and widely used procedure to achieve this target:

Distance matrix **D** is assumed Euclidian with elements D_{ij} representing the difference between vectors $\mathbf{x_i}$ and $\mathbf{x_j}$. Moreover, coordinates $\mathbf{y_1},\ldots,\mathbf{y_n}$ are obtained by projecting $\mathbf{x_1},\ldots,\mathbf{x_n}$ on a k-dimensional subspace. The distance between vectors $\mathbf{y_i}$ and $\mathbf{y_j}$ is denoted by δ_{ij}.

---

[60] Note: If we transform a gravitational equation with exponential error terms conforming to (2), the transformation into (10) yields a rather similar result, if the condition of uncorrelated error terms is met. The transformation including error terms would yield:

$$T_{ii}T_{jj} \big/ T_{ij}^2 = D_{ij}^* \ \mathrm{Exp}(\varepsilon_{ii})\ \mathrm{Exp}(\varepsilon_{jj}) \big/ \mathrm{Exp}(\varepsilon_{ij})\ \mathrm{Exp}(\varepsilon_{ji}) = D_{ij}^*\ \mathrm{Exp}(\varepsilon_{ii} + \varepsilon_{jj} - 2\varepsilon_{ij}).$$

Although $E(\varepsilon_{ii} + \varepsilon_{jj} - 2\varepsilon_{ij}) = 0$, the variance of this error term amplifies to $\sigma_{ii}^2 + \sigma_{jj}^2 + 4\ \sigma_{ij}^2 = 6\sigma^2$.

(11) $D_{ij} = \left\| \mathbf{x_i} - \mathbf{x_j} \right\|, \ \mathbf{D} = \{D_{ij}\}, \ \delta_{ij} = \left\| \mathbf{y_i} - \mathbf{y_j} \right\|$

Then the subspace minimising the expression in (12) is spanned by the eigenvectors corresponding to the k largest eigenvalues of **D**.

(12) $\sum_{ij} \left(\delta_{ij}^2 - D_{ij}^2 \right)$

Mardia/Kent/Bibby (1995, pp. 397-400) show that if we do not know the coordinate vectors x_i, they may easily be reconstructed if **D** is Euclidian.[61] The procedure works as follows: First, denote by **A** the n x n matrix with elements $a_{ij} = -\frac{1}{2} D_{ij}^2$. It is possible to construct its centered inner product matrix **B** by the matrix **H**, a unity matrix minus a n x n matrix with elements 1/n.

(13)
$$\begin{aligned} \mathbf{H} &= \mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^{\mathbf{T}} \\ \mathbf{B} &= \mathbf{HAH} = \mathbf{P}\Lambda\mathbf{P} \end{aligned}$$

The eigenvector decomposition of **B** (see (13)) yields a diagonal matrix of eigenvalues Λ and a matrix of eigenvectors **P**, which serve to construct the coordinate matrix **X**. Its columns $\mathbf{x_i}$ are obtained by (14), where $\mathbf{p_i}$ denotes eigenvector i, and λ_i the corresponding eigenvalue.

(14) $\mathbf{x_i} = \operatorname{sign}(\lambda_i) \sqrt{|\lambda_i|}\, \mathbf{p_i}$

According to (11), the $\mathbf{x_i}$ attributed to the two largest eigenvalues of **B** are the closest approximation to the distances of the “true” n-dimensional distance matrix **D*** feasible in two-dimensional space. Out of these two coordinate vectors we are able to create a map reflecting the coordinates attached the implied distance matrix **D***.

Figure 4 depicts such a map based on the EU-15 implied distance matrix, which in turn is derived from the EU-15 cooperation matrix. The implied distance matrix is displayed in Table A. 5 in the appendix. Total (logarithmic) cooperation intensity is reflected by the size of the nodes. The locations attributed to the member states reflect actual cooperation intensity very well.[62]

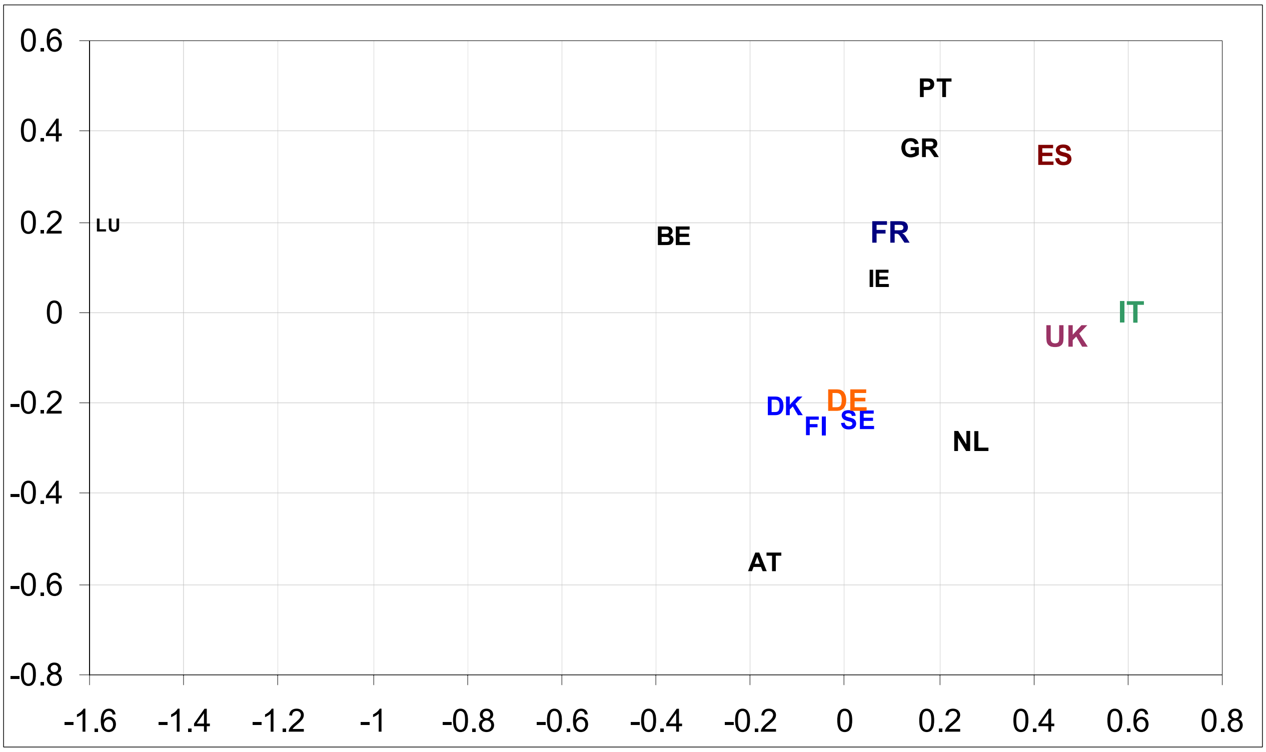
---

[61] For that purpose, the m dimensions of **D** have to be determined. In our case, the dimensions of matrix **D***, respectively its rank, are equal to n, the number of nodes. Therefore we skip reference to the case of m<n.

[62] The only exception is the case of Austria, whose respective coordinate in the third principal component differs considerably from the others. Thus its three-dimensional position may be imagined above the plane formed by the other 14 points.

The resulting map facilitates to draw insights from the cooperation matrix: The central part of the cooperation space (near to coordinates {0,0}) is dominated by the three countries most involved in FP cooperation: Germany, France and the UK. Furthermore, Ireland and Belgium are located in similar proximity to the origin.

Figure 4: Implied distances for EU-15 cooperation in the FP4[63]



The duchy of Luxembourg is positioned particularly far from the other 14 member states, which are roughly oriented along the vertical dimension (partly due to the specific position of Luxembourg).[64] By intuition, one may identify a North-South or Germanic-Romanic divide in this vertical dimension. While Germanic-speaking countries (along with Finland) cluster the lower part, countries adherent to the Romanic language group (and Greece) figure at the upper-right corner of the graph. The "central" member states France and UK may be interpreted as "connecting" nodes being equally involved with both sides.

---

[63] Two-dimensional scaling of the implied distance matrix derived from the cooperation matrix in Table 3 (log-transformed). Colours highlight Germany, Spain, France, Italy, the UK and the Scandinavian countries. The size of points is relative to the (logarithmic) total cooperation per node.

[64] It has to be noted that two-dimensional scaling on data excluding Luxembourg would result into about the same picture (not displayed) – the only major difference would be a more accurate positioning of Austria.

The hypothesis is confirmed if one considers Table 6: On average, the countries Germany, France and Britain make up for 45% of collaborative links of the "smaller" member states. However, the intensity of cooperation with each of those three varies considerably among the smaller countries. Countries with a considerable part of the population adherent to the Romanic language group over-proportionally cooperate with France, while the UK and Germany are the preferred partners of the Netherlands, Ireland, Austria and the Nordic countries. Of those two, the UK is slightly more linked up with "Romanic" countries, which as well is reflected by its positioning.

Table 6: Intensity of collaboration with Germany, France and the UK

	AT	BE	DK	ES	FI	GR
1) Total cooperative links	10,167	21,263	14,285	32,522	13,177	18,007
2) Links with DE, FR and UK	4,610	10,068	6,430	14,062	5,688	7,492
3) Links with DE in % of 2)	47.8%	33.3%	35.3%	31.6%	39.4%	32.0%
4) Links with FR in % of 2)	23.9%	33.7%	25.2%	33.7%	25.0%	29.9%
5) Links with UK in % of 2)	28.3%	33.0%	39.5%	34.6%	35.6%	38.1%

	IE	IT	LU	NL	PT	SE
1) Total cooperative links	9,140	45,435	831	33,168	13,114	20,806
2) Links with DE, FR and UK	4,320	21,470	340	15,514	5,543	9,768
3) Links with DE in % of 2)	25.8%	33.2%	41.5%	36.3%	30.5%	36.4%
4) Links with FR in % of 2)	25.7%	33.6%	34.1%	25.9%	30.8%	25.8%
5) Links with UK in % of 2)	48.5%	33.2%	24.4%	37.8%	38.7%	37.8%

In addition, it may be noted that Nordic member states cluster closely together, while their proximity to Germany is due to two-dimensional projection – Austria is actually nearer to Germany than the graph may suggest.

The equivalent positioning of the (log-transformed) NUTS-1 cooperation matrix is presented in Figure 5 and Figure 6.

Figure 5 illustrates that about half of the regions crowd a central area (encircled by the dashed line), while many nodes are located fairly distant from the centre. Specifically, this concerns Eastern German *Länder* and the poorer Spanish and Greek regions.[65] Several of the more important regions (i.e. those with large total cooperation numbers) rank as well

[65] To the most peripheral regions belong DE8, DED, DED, DEG (but not DE4/Brandenburg and DE3/Berlin); ES1, ES4, ES6 and ES7 (but not the corridor from Barcelona to Madrid ES2, ES3 and ES5), as well as GR1 and GR2.

outside the central area, most notably NL4, ES5 and a considerable share of important French and British regions.

Regarding the general positioning of countries, the Germanic-Romanic divide somehow reappears: Although German regions are sometimes located fairly distant from each other, they (along with Austrian regions) appear stretched along an area parallel to the 45° degree line, but shifted to the left. Belgian, Dutch and Nordic regions seem scattered among them, albeit in a position much nearer to the centre, respectively to the 45° line. In contrast, Italian, Greek and (to a lesser extent) Spanish regions seem to be oriented along an axis radiating from the centre to the lower right corner of Figure 5. British and French regions are scattered more evenly within and around the centre, although both appear to biased via the area right of the 45° line. Within the “Germanic” upper left part of Figure 5, the small cluster of UKG and the neighbouring, French-speaking regions BE3 and FR3 strikes the eye. As a general indicator, drawing a line from the coordinate pair (-1.3, -1.3) through the origin implies the entire set of Spanish, Greek, Italian regions and Portugal to be found below this line, along with a dominant share of French and British regions. Of the remaining countries, only DE2, DEB and NL4 are located below this line, albeit very near to it. In contrast, the only “Romanic” regions to venture the upper-left part are the French-speaking regions FR1, FR2, FR3, FR7 and BE3 (as well as the “mixed” regions BE1 and LU0). Apart from UKG, British regions above the separating line figure only within the very central region.

Figure 5: Implied distances for EU-15 NUTS-1 regions in the FP4[66]

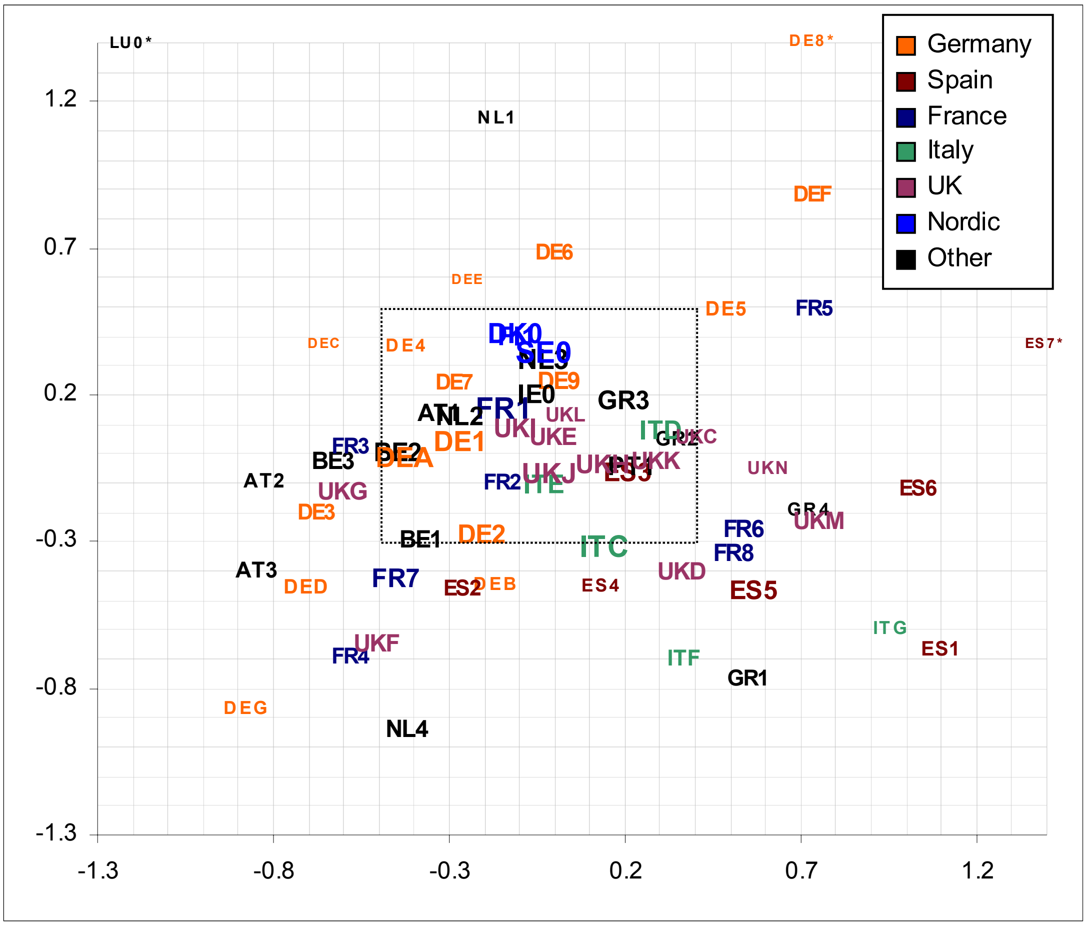
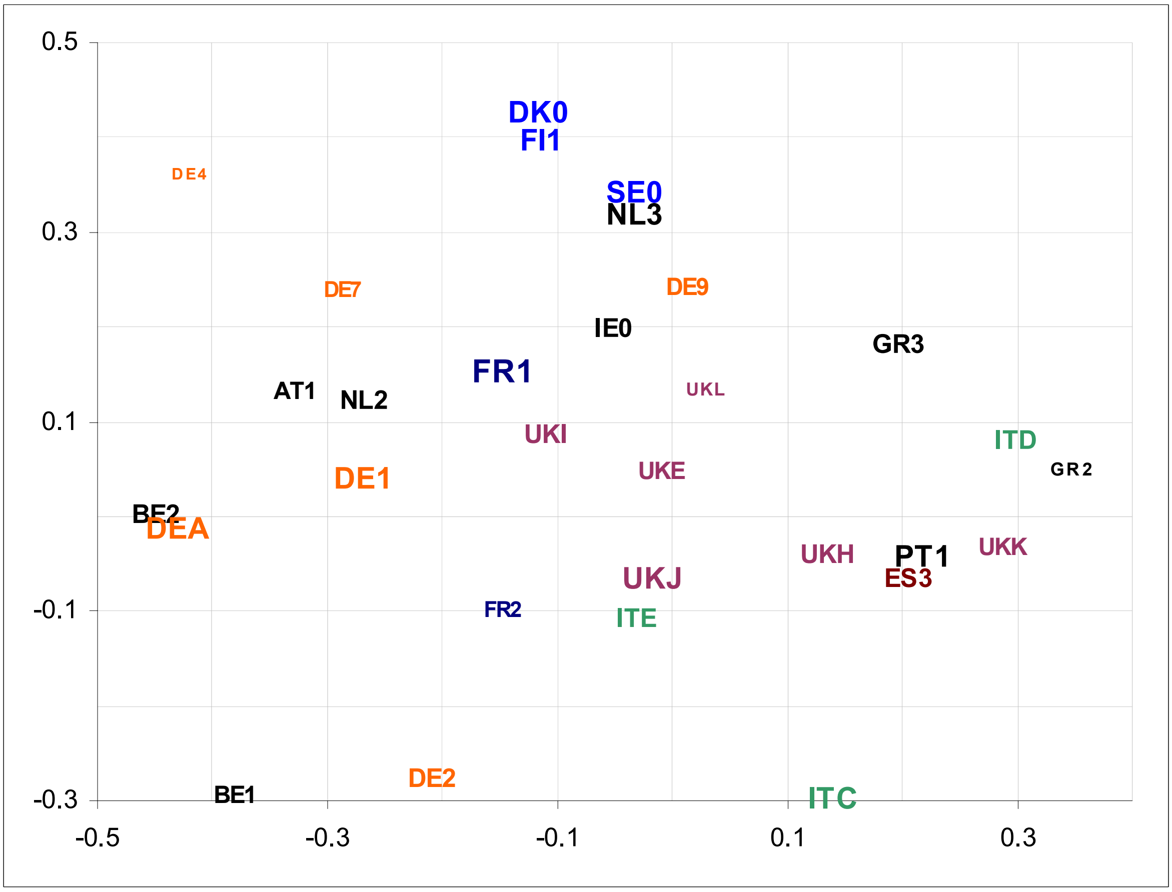


Figure 6 displays the regions with the least total (i.e. row sums) of implied distances, as demonstrated in Table A.4 in the appendix. Nodes are scattered around FR1 (Paris/Île-de-France), UKI and UKJ (London and South East) and DE1 (Baden-Württemberg). Ireland (IE0), mainland Portugal (PT1) and the Nordic countries (DK0, FI1 and SE0) figure as well in central positions – although one has to keep in mind that these five NUTS-1 regions each comprise a whole member state.[67] If they were as well divided into central and peripheral regions, they would be more comparable to Austria or Greece in this sample. Moreover, many regions representing the economic core of their respective countries are prominently located: NL3, DEA and the North Italian ITC, ITD and ITE rank among the most active contributors. Along with NL3, AT1, BE1, ES3 and GR3 represent capital regions, flanked by BE2, GR2 and NL2. While the largest German *Länder* scatter the left part of Figure 6, a cluster of Southern English and Welsh regions catches the eye in the centre.

[66] Nodes denoted by asterisks actually are located outside the Figure's boundaries, but were shifted towards the centre for illustration purposes.

[67] PT1 is Portugal less the Açores and Madeira, FI1 is Finland less the Åland Islands.

Figure 6: Implied distances for EU-15 NUTS-1 regions in the FP4 (central part)



The presented distribution leads to the supposition that regions representing the economic cores of the corresponding member states constitute the centre of the map. This behaviour is embodied in the implied distance matrix, since a large number of intra-regional cooperation links in node i is inversely related to its implied distance to other nodes j – and intra-regional collaboration of i is well correlated with the total number of external links.[68] Thus it comes as no surprise that regions boasting high "mass" indicators are rather more positioned in the centre of the cooperation spectre. However, one might ask why regions such as DE9 (Lower Saxony) with a total cooperation number of 6,060 are much more centrally located than, say, UKM (Scotland) with 6,211 collaborative links. The reason for this may partly lie in "true" mass measures apart from total cooperation, and partly in exogenous distances – a point further enquired during the next sections.

[68] Comparing the 68 main diagonal elements of the NUTS-1 cooperation matrix with its 68 row sums less the diagonal elements yields an R^2 of 0.882.

5.1.3 Node-specific Propensity to Collaborate – Implied "Masses"

While distance indicators reflect inter-nodal relationship, the gravity model approach calls as well for node-specific factors determining the general propensity to collaborate – the so-called "masses", in Newtonian terminology. As already mentioned in section 5.1.1, an n x 1 mass vector may already explain a large part of variance in the collaboration matrix. If the gravity model holds, and if it is combined with a "true" distance matrix, a single vector of masses (times a constant) should explain cooperation (along with an error term).

By using implied distance and by taking advantage of the collaboration matrix's convenient properties, we are able to derive an n x 1 vector depicting implied mass effects. For this purpose, we perform an algebraic transformation based on an idea by Shen (1999, pp. 215-218).[69]

We depart from a deterministic model structure similar to (2):

(15) $$T_{ij} = \theta \frac{M_i^\alpha M_j^\alpha}{\mathrm{f}(D_{ij})} \qquad \text{where } T_{ij} \equiv T_{ji}$$

Furthermore, we assume the distance matrix (with elements $\mathrm{f}(D_{ii})$) to be symmetric. Consequently, the model structure implies a symmetric collaboration matrix with elements $T_{ij}=T_{ji}$.

Due to symmetry, the inter-nodal relationships in the collaboration matrix are fully described by the n(n+1)/2 equations in its lower triangular matrix (16).

(16) $$\begin{array}{llll} M_1^\alpha M_2^\alpha = T_{12}\,\mathrm{f}(D_{12})\,\frac{1}{\theta} & \ddots & & \\ M_1^\alpha M_3^\alpha = T_{13}\,\mathrm{f}(D_{13})\,\frac{1}{\theta} & M_2^\alpha M_3^\alpha = T_{23}\,\mathrm{f}(D_{23})\,\frac{1}{\theta} & \ddots & \\ \vdots & \vdots & & \\ M_1^\alpha M_n^\alpha = T_{1n}\,\mathrm{f}(D_{1n})\,\frac{1}{\theta} & M_2^\alpha M_n^\alpha = T_{2n}\,\mathrm{f}(D_{2n})\,\frac{1}{\theta} & \cdots & M_{n-1}^\alpha M_n^\alpha = T_{(n-1),n}\,\mathrm{f}(D_{(n-1),n})\,\frac{1}{\theta} \end{array}$$

Multiplying all equations of (16) where M_i takes part results in the expressions in (17):

---

[69] Shen (1998) algebraically transforms the expression in (15), with α set to 1, in order to display unknown masses M_i as a function of known interaction T_{ij} and known distance $\mathrm{f}(D_{ij})$. His result should be equal to (20) in case a=1, but due to a error in deduction (Shen 1998, p. 217), his equivalent of formula (20) is different from the ours – albeit the relationship between his and our formula is linear.

(17)
$$M_1^{\alpha(n-2)}\prod_{j=1}^{n}M_j^{\alpha} = \left(\frac{1}{\theta}\right)^{n-1}\prod_{j=2}^{n}T_{1j}\ \mathrm{f}(D_{1j})$$
$$M_2^{\alpha(n-2)}\prod_{j=1}^{n}M_j^{\alpha} = \left(\frac{1}{\theta}\right)^{n-1}T_{12}\ \mathrm{f}(D_{12})\prod_{j=3}^{n}T_{2j}\ \mathrm{f}(D_{2j}) = \left(T_{22}\ \mathrm{f}(D_{22})\right)^{-1}\prod_{j=1}^{n}T_{2j}\ \mathrm{f}(D_{2j})$$
$$\vdots \qquad \vdots$$
$$M_i^{\alpha(n-2)}\prod_{j=1}^{n}M_j^{\alpha} = \left(\frac{1}{\theta}\right)^{n-1}\prod_{j=1}^{i-1}T_{ij}\ \mathrm{f}(D_{ij})\prod_{j=i+1}^{n}T_{ij}\ \mathrm{f}(D_{ij}) = \left(T_{ii}\ \mathrm{f}(D_{jj})\right)^{-1}\prod_{j=1}^{n}T_{ij}\ \mathrm{f}(D_{ij})$$

A node-specific mass can hence be described as in (18):

(18)
$$M_1^{\alpha} = \sqrt[n-2]{\left(\theta^{n-1}\prod_j M_j\right)^{-1}\prod_{j=2}^{n}T_{1j}\ \mathrm{f}(D_{1j})} = \sqrt[n-2]{\left(\prod_j M_j^{\alpha}\right)^{-1}\theta^{-(n-1)}\frac{1}{T_{11}\ \mathrm{f}(D_{11})}\prod_{j=2}^{n}T_{1j}\ \mathrm{f}(D_{1j})}$$
$$M_i^{\alpha} = \sqrt[n-2]{\left(\theta^{n-1}\prod_j M_j\right)^{-1}\prod_{j=1}^{i-1}T_{ij}\ \mathrm{f}(D_{ij})\prod_{j=i+1}^{n}T_{ij}\ \mathrm{f}(D_{ij})} = \sqrt[n-2]{\left(\prod_j M_j^{\alpha}\right)^{-1}\theta^{-(n-1)}\frac{1}{T_{ii}\ \mathrm{f}(D_{ii})}\prod_{j=1}^{n}T_{ij}\ \mathrm{f}(D_{ij})}$$

The product of all masses being part of (18) is obtained by multiplying all n(n+1)/2 equations in **(16**) – the corresponding result is displayed in (19):

(19) $$\left(M_1M_2....M_{n-1}M_n\right)^{\alpha(n-1)} = \prod_{j=1}^{n}M_i^{\alpha(n-1)} = \left(\frac{1}{\theta}\right)^{n(n-1)/2}\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}T_{ij}\ \mathrm{f}(D_{ij})$$

Substituting the expression ΠM^{α} in (18) by the (n-1)-th root of (19) leads to (20):

(20) $$M_i^{\alpha} = \sqrt[n-2]{\frac{\theta^{-(n-1)}\dfrac{1}{T_{ii}\ \mathrm{f}(D_{ii})}\displaystyle\prod_{j=2}^{n}T_{ij}\ \mathrm{f}(D_{ij})}{\left(\dfrac{1}{\theta}\right)^{-n/2}\left(\displaystyle\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}T_{ij}\ \mathrm{f}(D_{ij})\right)^{1/n-1}}}$$

So, if a distance matrix with elements $\mathrm{f}(D_{ii})$ is known, it is straightforward to calculate a mass vector with elements M_i. Although we do not dispose of the "real" distance matrix, we assume the properties needed for implied distances to hold. The assumption of internal distances $\mathrm{f}(D_{ii})$ to be equal over all nodes enables us to simplify (18) to (21).

$$\mathrm{f}(D_{ij}) \approx D_{ij}^{*} = \frac{\sqrt{T_{ii}T_{jj}}}{T_{ij}} \qquad \text{where } D_{ii}^{*} \equiv 1 \quad \forall i$$

(21) $$M_i^{\alpha} = \sqrt[n-2]{\left(\prod_j M_j^{\alpha}\right)^{-1}\theta^{-(n-1)}\frac{1}{T_{ii}}\prod_{j=1}^{n}T_{ij}\frac{\sqrt{T_{ii}T_{jj}}}{T_{ij}}} = \sqrt[n-2]{\left(\prod_j M_j^{\alpha}\right)^{-1}\theta^{-(n-1)}T_{ii}^{\frac{n-2}{2}}\prod_{j=1}^{n}\sqrt{T_{ii}}}$$

Similarly, (19) can be transformed into (22):

(22) $$\prod_{j=1}^{n} M_i^{\alpha} = \left(\frac{1}{\theta}\right)^{n/2} \sqrt[n-1]{\prod_{i=1}^{n-1}\prod_{j=i+1}^{n} T_{ij}\frac{\sqrt{T_{ii}T_{jj}}}{T_{ij}}} = \left(\frac{1}{\theta}\right)^{n/2} \sqrt[n-1]{\prod_{i=1}^{n} T_{jj}^{(n-1)/2}} = \left(\frac{1}{\theta}\right)^{n/2} \prod_{i=1}^{n} T_{jj}^{1/2}$$

Combining (21) and (22) and further transforming yields the expression in (23):

(23) $$M_i^{\alpha} = \sqrt[n-2]{\frac{\theta^{-(n-1)} T_{ii}^{\frac{n-2}{2}} \prod_{j=1}^{n}\sqrt{T_{ii}}}{\theta^{-n/2}\prod_{i=1}^{n} T_{jj}^{1/2}}} = T_{ii}^{\frac{1}{2}}\theta^{-\frac{1}{2}}$$

By definition, the resulting matrix vector combined with distance should explain nearly the total of cooperation matrix variance. However, without prior knowledge of θ, it is only possible to determine the relative rather than the absolute sizes of the mass functions M_i^{α}. For the purpose of this study, absolute "mass" sizes are irrelevant, since the resulting values would have no meaning. Rather, we are interested in the relative distribution of mass functions. For this reason, we define mass functions linearly scaled by the factor $\theta^{½}$ as in (24).[70]

(24) $$M_i^{*} \equiv M_i^{\alpha}\theta^{\frac{1}{2}} = T_{ii}^{\frac{1}{2}}$$

A recombination of implied distances and implied masses as in (25) yields exactly the empirical collaboration matrix. The collaboration matrix **T** is thus explained by the outer product of implied mass vector **m*** in combination with implied distance matrix **D***.

(25) $$T_{ij} = \frac{M_i^{*}M_j^{*}}{D_{ij}^{*}} \qquad \forall\, i, j \text{ where } M_i^{*} \equiv M_i^{\alpha}\theta^{\frac{1}{2}},\ D_{ij}^{*} \equiv \mathrm{f}\left(D_{ij}\right) \text{ if } \mathrm{f}(D_{ii}) = 1$$

Besides, the approach of implied masses is akin to drawing the eigenvector corresponding to the largest eigenvalue of a symmetric matrix whose elements consist of T_{ij} multiplied by D_{ij} (for all i,j). In fact, the resulting eigenvector boasts a correlation coefficient of 0.99 to implied masses and is equivalent to a normalised and centralised implied mass vector. Nevertheless,

---

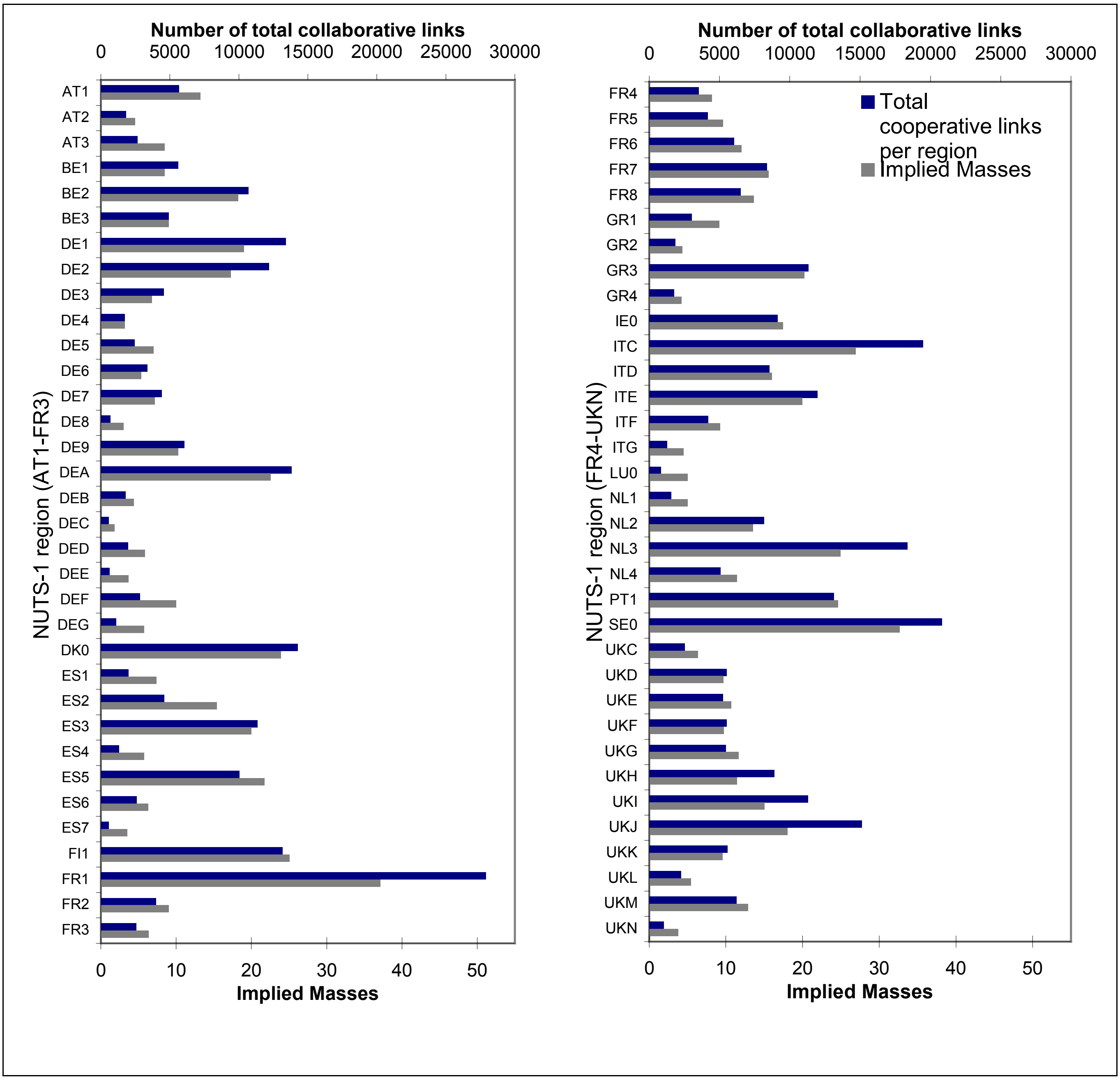
70 Lifting the assumption of $f(D_{ii})=1$ would result into implied distance $D_{ij}^{*} = T_{ii}T_{jj}\,\mathrm{f}(D_{ii})\mathrm{f}(D_{jj}) / T_{ij}$

Consequently the expression in (23) would become $M_i^{\alpha} = T_{ii}^{\frac{1}{2}}\theta^{-\frac{1}{2}}\,\mathrm{f}\left(D_{ij}\right)^{\frac{1}{2}}$. Hence implied mass M_i^{*} would be defined as $M_i^{*} \equiv M_i^{\alpha}\theta^{\frac{1}{2}} = T_{ii}^{\frac{1}{2}}\,\mathrm{f}\left(D_{ij}\right)^{\frac{1}{2}}$. This is useful if one has a guess of internal distance per node.

we stay with implied masses, since their definition facilitates intuitive understanding and economic interpretation.

Figure 7 displays the implied masses of the NUTS-1 FP4 collaboration matrix in comparison with total collaboration links per region. Overall, implied masses vary less than total cooperation numbers.[71]

Figure 7: Implied masses versus the number of collaborative links (row sums)



[71] The standard deviation of implied masses divided by its average amount to 0.64, while this figure numbers 0.87 for total collaboration links.

Since implied masses adjust for distances, this discrepancy simply reflects the core-periphery situation of nodes. This interpretation is supported by a correlation coefficient of 0.52 between cooperation-implied mass differences and the (logarithmic) row sums of the implied distance matrix. Nevertheless, the general impression changes just slightly: Economic core regions exhibit the largest values. However, the picture is somewhat altered at less active nodes. The regions ES2 or DEG, for instance, show much improved numbers, while nodes like DE7 apparently lose in importance compared to total collaboration intensity. It is also instructive to pay attention to the intra-national distribution of implied masses: while England/Scotland exhibit a rather even distribution, France's and Greece's capitals seem to attract activity over-proportionally. Furthermore, Italian and Spanish regions exhibit signs of the familiar north-south disparities… Further analysis requires implied masses to be set in relation to other indicators – section 5.3 will enquire their distribution in more detail.

5.2 Empirical Distance Measures

In order to interpret FP cooperation as a gravity pattern, a matrix of exogenous distances is required in order to account for mass-adjusted inter-node relationship. With respect to regional data, the obvious candidate is geographic distance between nodes: In analogy to Newton's theory and to numerous regional science models, rising geographic distance between nodes would decrease interaction in a continuous manner. However, research collaboration may not (or not only) depend of geographic distance, since this kind of interaction may rather be shaped by similarities or dissimilarities concerning research topics, language, other interaction standards or simply the frequency of personal contacts (see section 2.2 on a discussion of literature on that topic).

In order to identify a generic, exogenous distance indicator determining FP collaboration, we therefore examine several factors of potential influence. Unfortunately, useable data on the regional level is hard to find, therefore we were not able to compile more than eight prospective distance matrices: One of them is geographic distance, while three represent cultural and language ties in accordance with section 2.2.4. Unfortunately, only one of them embodies distances on a regional level, while the other two represent differences only between nation states. Furthermore, we introduce four measures indicating structural dissimilarity – following the argumentation laid out in section 2.2.5.

This section will illustrate the compilation of the four distance matrices, and will examine their resemblance to the implied distance as described on pp.87-99.

5.2.1 Geographic Great Circle Distance

The most obvious distance matrix is one of the most straightforward to compile: The distance between two geographic points is sufficiently defined, and data on the geographic coordinates of NUTS-1 regions is relatively easy to obtain.

In order to calculate distance between nodes, the geographic centre (or centre of gravity) of the land surface of region i provides for its coordinates.[72] Taking the focal point of regional population or firm distribution may be a superior approach, but we lack data on these figures. However, the lower the regional des-aggregation, the less important the difference between the geographic and the economic centre becomes. On the NUTS-1 level we judge discrepancy between those two central points to be relatively minor compared to their distance to other regions.[73]

Out of the thereby obtained coordinates it is straightforward to calculate great circle distances in kilometres (i.e. the surface distance between to two points assuming that earth is a perfect sphere with a radius of 6,378 km).

For the effects described in section 2.2.3, we expect geographic distance to exert negative influence on collaboration intensity, since we assume that the likelihood of personal interaction (a pre-constitute to collaboration) decreases with distance. Yet the relationship between the frequency of personal contacts and great circle distance may hold only for shorter distances, which are overcome by land transport. Trips over longer distances depend almost solely on air travel: And concerning the short distances in Europe, this characteristic renders the associated transport cost and time almost equal for the entire continent, since the fixed cost and time effects with departure and arrival mostly outweigh the variable costs

[72] We derived the geographic centres from a routine implemented in the ArcView GIS software package.

[73] The region FI1 (Manner-Suomi / Finland), for instance, is considered a case where the distance between geographic and economic centre is considerable: Population and economic power are concentrated at the Gulf of Finland, while the vastness of the country's northern regions induce the geographic centre to be located considerably farther in the North. E.g. the geographic centre of FI1 and the Helsinki region FI181 are more than 400 km apart. Nevertheless, this roughness is outweighed by Finland's distance to other EU-15 regions: Only one region's geographic centre (that of Sweden) is nearer to Finland than 1,900 km. The same reasoning applies to other vast regions whose economic centre is situated at the periphery of their land surface (e.g. UKL / Wales). In most cases, these large regions are located at the very periphery of the European Union, thus a potential mismatch between geographic and economic centre may be outweighed by considerable distance to all other nodes.

of a trip. Therefore one may ask whether the impact of geographic distance becomes marginal if it surpasses a certain “air trip” threshold.

However, since inter-regional collaboration is likely to be linked to the frequency of personal contacts between regions, there may still be an effect of geographic distance beyond the “air trip threshold”: Many surveys of international trade identify an effect of geographic distance on trade flows (compare Porojan 2000), since the marginal costs of e.g. shipping goods by land do not decrease stepwise as in air travel. And trade is one cause of personal interaction between two regions. Culture may be another factor – and cultural similarities and disparities were shaped by historical diffusion, which in turn followed geographic circumstances. Several other soft factors come to mind raising doubts on whether an “air trip” threshold may be significant. (In order to determine its relevance, we assessed the impact of an adjusted distance measure, but obtained no meaningful results).

5.2.2 Inter-node language differences

As already mentioned in section 2.2.4, common language between nations or regions facilitates interaction considerably. Trade economics, for instance, provides us with an extensive body of literature examining the effects of language in gravity models of international import and export flows. The predominant part of these studies includes dummy variables taking the value 1 if the regions examined share a common native language and zero otherwise. Rather than using simple dummies, recent research has examined proportions of native/bilingual speakers (Mélitz 2002). Concerning our regional data set, such an approach would capture the relationship between the Walloon region and France, for instance. However, solely including native language data would not reflect the sharing of common lingue francae, i.e. English, as well French in some countries, Swedish in the Finnish-Swedish relationship, etc. Fortunately, the European Commission (2002) provides us with data on the principal languages[74] studied in schools at “ISCED level 3” (upper secondary) education level for most NUTS-1 regions.

By adding the numbers of native speakers we are able to attribute the number of speakers of each major language to each region in our data set. The thereby obtained data lists may be interpreted as vectors in a space dimensioned by languages (each language represents one dimension). The degree of overlapping between two of those vectors is expressed by their angle to each other: By simple algebraic reasoning we know that the inner product of two vectors of norm (length) 1 is equal to the cosine of their angle (as displayed in (26)). The

[74] These encompass the eleven official languages of the EU-15 and Arabic, Chinese, Japanese and Russian.

farther two vectors are apart, the lower their inner product (the cosine) and hence the greater their angle. Thus it is straightforward to normalise the language vectors attributed to NUTS-1 regions and calculate a matrix of bilateral angles.

(26) $$\cos \alpha_{IJ} = \frac{\mathbf{m}_\mathbf{I}' \mathbf{m}_\mathbf{J}}{\|\mathbf{m}_\mathbf{I}\| \|\mathbf{m}_\mathbf{J}\|} = \mathbf{m}_\mathbf{I}' \mathbf{m}_\mathbf{J} \left(\sum_i m_{I,i}^2 \sum_i m_{J,i}^2 \right)^{-0.5}$$

These angles reflect the distance in language orientation between two nodes – i.e. the more priorities of languages study differ, the less voluminous is the common body of language knowledge and the higher the figure in the language matrix. If the reasoning behind these figures is relevant to FP collaboration, we expect it to exert negative influence on cooperation intensity between regions.

Numbers on language education, however, have some drawbacks in representing language differences, since they may not accurately reflect the languages actually spoken in a region.[75] In order to obtain a closer measure of language skills, we draw on annual survey data published by European Commission (2001a). The survey details the respective percentages of speakers of foreign tongues in EU-15 member states, albeit only at the national level. We combined this aggregated data with information on countrywide shares of native speakers and apply the method as described by (26). Consequently we obtain a matrix of nation-to-nation language vector angles in 11-dimensional space. Using the strong assumption of language shares to be equal across all over regions in a country, we attribute the same figures to NUTS-1 regions (this implies a value of 0 for intra-national angles).

Both language distance matrices are compiled as a proxy for cultural relatedness among researchers and their organisations across regions. However, provided this cultural effect exists, it is questionable whether those figures provide an accurate measure: First, both sets of language data relate to the general population, rather than to the specifics of the research community. Second, angles between vectors describe bilateral proximity in relation to other vectors, but do not contain information on each vector's length (i.e. the share of inhabitants speaking a foreign language). Third, languages are weighted equally regardless of their importance in international communication.

[75] First, the quality and intensity of language education must at minimum enable students to communicate in the respective language – and these requirements in education may differ with respect to the language learnt. Second, Eurostat schooling data describes school teaching during the 1990s, while the labour force may dispose of different language skills (the most extreme case would be the shift from Russian to English in the new German Länder).

Yet the likelihood of cooperation between researchers is strongly related to prior personal interaction (as stated in section 2.1.6), and this interaction in turn depends on the frequency of interaction between those researchers' groups, institutions or regions. Therefore, the more interaction takes place between regions as a whole, the more intense collaboration may be reckoned. But it remains questionable whether the presented indicators of language differences accurately reflect inter-regional interaction potential.

5.2.3 Inter-node cultural differences

By the same reasoning as above, disparities in cultural attributes could constitute barriers to interaction. In that respect, we draw on country data by Hofstede (1980) representing cultural "dimensions". Although the measuring of cultural characteristics in general and Hofstede's methods in particular have given rise to fierce academic debate, his data is still referred to as the standard in the culture-economics nexus.

Hofstede (1980) provides data on the four dimensions "power distance", "individualism", "uncertainty avoidance" and "masculinity-femininity" for each of the EU-15 member states. We assume the figures to be equal for the total each nation's regions, and compile angles between the resulting vectors by the method described in (26). Since increased cultural distance should lead to less frequent interaction between countries, we except the resulting cultural distance matrix to induce negative effects on collaboration intensity, provided such an effect exists. As with language data, we have doubts on its capability to indicate actual cultural barriers.

5.2.4 Disparities in Firm-Size Structure

From section 2.1.5 we know that firm size plays a central role in most publications on research collaboration. But not only their absolute, but also their relative size to each other may promote or impede cooperation. In order to measure this relative effect, we draw on agglomerated data from the Amadeus database (Amadeus 2003) on European enterprises. In particular Amadeus provides with the number of firms per region and divided into one of four categories depicting the number of employees.[76] Again we interpret these four categories per region as 68 4x1 vectors and calculate the angles between those according to

[76] Note: The staff number dimension separates into firms with less than 200, 200 to 500, 500 to 1,000 and more than 1,000 employees. The data concerned does not include all entries in the database, since coverage may vary from country to country. Instead, we only include the firms which belong to the European Top 1,500,000 by Amadeus definition. Unfortunately, we were not able to determine the respective figures for the 1990s, therefore Amadeus (2003) data included stems from 2003.

(26). By the reasoning taken from the literature review, we expect inter-regional disparities in firm size distribution (i.e. larger angles) to induce a negative effect on collaboration.

5.2.5 Inter-Industry Disparities in Innovation Structure

Apart from “basic” distance measures, symmetry/asymmetry in the regional structures of the research sector might promote or impede cooperation between nodes. Section 5.3 describes Eurostat research data later used to model mass effects. From these data, we are able to construct three sector-specific distance measures, one among them reflecting disparities in patent structure:

Eurostat provides fairly detailed data on *patent applications* at the European Patent Office, and attributes these applications to eight distinct sectors.[77] For many regions and sectors, these data are too infrequent and discontinuous to be used in least squares estimation. However, compiling the angle between those 8x1 vectors (as in (26)) yields an easy-to-use, continuous distribution. The resulting angles should roughly represent the degree of disparity between regional effectiveness in innovation. This mainly relates to the private sector (since patents are mostly requested by profit-oriented entities) and represents innovations ready for use. It is questionable whether “patent structure distance” may be linked to FP collaboration, since the scheme is dominated by non-profit institutions and concentrates on generic research. Moreover, it is unclear how this indicator may affect collaboration from the theoretical point of view: If the data represents asymmetries in the focus of private R&D, they are subject to the reasoning presented on sector asymmetry in firm RJVs (see pp. 17-21). At the microeconomic level, Röller/Tombak/Siebert (1997) reckon sector asymmetry to encourage inter-firm collaboration, although they cannot corroborate their findings empirically. Navaretti et al. (2002), in contrast, provide convincing arguments for formal research collaboration to be promoted by sector symmetry of participants. A positive correlation between “patent structure distance” and implied distance would therefore support the hypothesis of Navaretti et al., while a negative relationship would point to the propositions by Röller/Tombak/Siebert.[78]

---

[77] The eight sectors for which patent application data is provided: A (human necessities), B (performing operations, transporting), C (chemistry, metallurgy), D (textiles, apparel), E (fixed constructions), F (engineering, lighting, heating), G (physics), H (electricity). Source: Eurostat (2003, Domain r_epa). See p. 109 for further discussion of patent application data.

[78] As depicted in Table 7, correlation between the patent structure distance indicator and implied distance is fairly positive, which would support the case of Navaretti et al. (2002).

5.2.6 Disparities in Research Sector Structure

In addition to patent applications, Eurostat delivers data on regional research expenditure and research personnel, both disaggregated into business, government and higher education sector. (For detailed discussion of the data, please refer to p. 108.) This provides us with two (one for expenditure and one for personnel) 3x1 vectors detailing the relative regional importance of those sectors. Conforming to (26), we compute the angles between the regional vectors, resulting into a “distance” matrix. A priori, it is difficult to assess how disparities in expenditure or staff vectors might affect research cooperation, if there is an impact at all. On the one hand, participations in an FP project may be promoted by asymmetries, if organisations in the respective regions perceive the scheme as an opportunity to overcome insufficient input by a specific regional research sector. On the other hand, empirical data on FP collaboration shows that education institutions (mainly universities) are concentrated in certain key action lines, while they are less present in production-oriented fields. Based thereupon, one might conclude that depending on a project’s nature, either public-public, or private-private cooperation dominates. This in turn would imply that organisations in regions with a strong private sector would be more inclined to cooperate with regions boasting important private sector research as well. This would constitute an analogy to Navaretti et al. (2002), who find increased collaboration between firms with additive R&D resources, but substitutable products. The resources and “products” of public research institutions may follow a similar pattern. Thus both kinds of impacts may be possible, although we tend to the follow the latter statement. This would also be in line with results delivered by Hussler (2003).

5.2.7 Correlation between Implied and Empirical Distance

In order to examine the explicative power of the geographic, linguistic and cultural distance and research sector measures introduced in this section, it is helpful to analyse their correlation with the implied distance matrix obtained from (10). More formally, we focus at the correlation between the series obtained by “stacking” the matrices: Since all of the matrices in question are symmetric, their upper triangular parts are redundant – thus we only consider the main diagonal and lower triangular parts. “Stacking” refers to taking the lower triangular observations of the columns of the respective matrix **D** (with n-i+1 observations in the i-th column) and compile a vector $\overline{D}$ containing the observations of all those columns in order of their matrix element index (compare (27)).

$$\textbf{(27)}\quad \mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1n} \\ D_{12} & D_{22} & \cdots & D_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ D_{1n} & D_{2n} & \cdots & D_{nn} \end{pmatrix} \Rightarrow \overline{D} = \begin{pmatrix} D_{11} \\ \vdots \\ D_{1n} \\ D_{22} \\ \vdots \\ D_{2n} \\ \vdots \\ D_{nn} \end{pmatrix}$$

The correlations between the stacked series and between their log-transformations are depicted in Table 7. It shows that correlations with log-transformed implied distances are considerably higher than those with not-transformed implied distances. Unsurprisingly, it seems that the linkage to research sector "distances" seems more pronounced than to other indicators. Of the remaining measures, geographic distance appears as the most correlated with implied distances while the language indicators and relative firm size structure are of lesser importance. Differences in cultural dimensions appear to bear no relationship to implied distances at all.

Table 7: Correlations between implied distance and empirical distance indicators

	Geographic distance	Spoken language	Language studied	Cultural dimensions	Firm size	Research expenditure	Research personnel	EPO patent applications
Identifier	*GEODIST*	*LANGSPOK*	*LANGSTUD*	*CULTDIM*	*INDSTRUCT*	*RDXDIFF*	*RSTAFFDIFF*	*PATSTRUCT*
Correlation with:	**Absolute values (not transformed)**							
Implied distance	0.160	0.133	0.072	-0.056	0.120	0.230	0.226	0.231
Log- implied distance	0.265	0.217	0.149	0.015	0.138	0.313	0.305	0.325
Jarque-Bera	291.8	803.9	236.2	35.3	1011.5	177.6	162.3	1.4
Prob. value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.506
Correlation with:	**Log-transformed**							
Implied Distance	0.164	0.129	0.076	-0.055	0.112	0.229	0.224	0.228
Log-implied distance	0.304	0.215	0.157	0.019	0.131	0.315	0.306	0.327
Jarque-Bera	19599.9	1056.0	335.2	60.9	705.8	81.3	85.3	30.9
Prob. value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Combining several of these distance indicators may even yield more explicative power than one single series. This purpose requires more sophisticated methods, which in turn depend on the distribution of the series examined. In particular, the requirement of normally distributed error terms in least squares regression analysis is more easily accomadated with normal distribution of exogenous and dependent series. Histograms of the series (not displayed), however, show that the distribution not-transformed series differs from normality: Table 7 depicts the Jarque-Bera value, a general statistic determining whether the hypothesis of normal distribution is likely. The higher its value, the lower its probability value and thus the less likely the data under question is normally distributed. This implies that most of the presented distance indictors are unlikely to follow normal distribution, but their log-transformed counterparts even less so. Nevertheless one has to bear in mind that the rejection of the normality hypothesis stems from the main diagonals of the empirical distance matrices, which exhibit similar and very low values by definition. Besides, it is hard to imagine geographic distance as being a stochastically distributed variable.

5.3 Empirical “Mass” Indicators

Since this study is characterised as exploratory, the primary goal was to gather as much as possible of regional data relevant to research. For this purpose, we rely mainly on data from Eurostat, predominantly from its regional Science and Research Indicators dataset (Eurostat 2003). Furthermore, general economic data such as the number of employees, etc. are introduced. This already large set of possibly explanatory factors is enhanced by data on the number of firms by size from Amadeus (2003). The following pages will first describe necessary data pre-processing and than provide a summary of the indicators used. Subsequently, a range of possible combinations of these indicators will be presented. A brief summary of possible dummy variables follows.

5.3.1 Data Pre-processing

Eurostat provides an extraordinary range of data on the state of research in NUTS-1 regions. These indicators belong to four broad categories: “Human resources in Science and Technology” (a human capital stock indicator), “Number of Researchers”, “Research Expenditure” and “Patents”. Each of those is further disaggregated by economic sector.

Unfortunately, many of those figures are updated only in irregular intervals. Moreover, these intervals differ by member state. This characteristic leaves us with a data set mainly consisting of huge gaps in annual data series.

For the analysis of collaboration in the FP4, the intention was to gather data from the time period 1994 to 1998. The year with most complete data on Research and Science data was 1996, therefore it was chosen as the basic year of reference. However 1996 data was still subject to considerable data gaps. In order to fill the empty observations of the more important indicators, we imputed available data from 1995, 1997 and 1998 (and even 2000 in the case of research staff numbers). Since the NUTS-1 structure was subject to reconfiguration in recent years, we then had to reshuffle NUTS-2 figures for the member states concerned.

The case was simpler for general statistic indicators, since complete series were available for 1996. Firm size data by Amadeus, though, was only delivered for the “last available year”. Since data was drawn in 2003, Amadeus figures are unlikely to represent the 1990s. Still, this approach to data consolidation seems justifiable, since apart from general statistics all of the data concerned are stock indicators.

5.3.2 Consolidated Mass Indicators

“**Human Resources in Science and Technology**” (HRST) constitutes one of the most comprehensive categories in the Eurostat dataset. A person is defined as HRST if she has either successfully completed third-level education in an S&T field of study or, if not formally qualified, is employed in an S&T occupation normally requiring this education (Eurostat 2003, Domain HRST). The notion of HRST comprises several sub-categories: “occupation” (HRSTO), persons employed in an S&T profession; “education” (HRSTE), persons having completed third-level S&T occupation; and “core” (HRSTC), people who both dispose of third-level S&T education and work in an S&T occupation.

Based on survey results, Eurostat provides regional stock data on the number of persons classified among the four HRST sub-categories. Apart from general HRST data, Eurostat provides sufficient regional data on the private and public service sector (NACE classification codes 50 to 93), as well as on “Knowledge-intensive Services” (a Eurostat definition comprising 14 two-digit NACE code categories).

Thereby we obtain twelve highly collinear mass vectors: HRST, HRSTO, HRSTE and HRSTC for the total of sectors, for services and for knowledge-intensive services. We certainly expect these variables to exert positive influence on FP collaboration – though a priori it is not sure which of the twelve vectors would most accurately reflect mass effects.

“**Research and Development Expenditure and Personnel**” (Eurostat 2003, Domain rd_ex_p) is as well based on surveys. Regional data on “R&D expenditure” details the intramural expenditure on R&D purposes, including both current and capital expenditure. Figures are provided in Euro, Purchasing Power Parities and percentage of GDP. Inter alia, “R&D Personnel” counts regional research staff in full-time equivalents, including persons employed in research and supporting operations.

Both expenditure and personnel data are broken down by economic sector: “Business enterprise sector” (BIZ) covers organisations involved into market production; “Government sector” (GOV) comprises virtually all public research efforts apart from higher education and public enterprises; and “Higher education sector” (EDU), including all kinds of post-secondary educational institutions and their subordinated research departments.[79] As with HRST, we expect broad expenditure and personnel figures to induce positive effects on R&D cooperation. A priori, it is not completely determined which implications to draw from an over-proportional representation of a particular sector. The sub-section 5.3.3 will summarise how ratios may be used to investigate such effects.

The third large Eurostat data domain, “**Patent Applications to the European Patent Office**”, provides detailed data on those applications by manufacturing sub-sector and NUTS regions. In contrast to previously mentioned Eurostat domains, patent data is based on actual filings rather than surveys and is virtually completely present for each year and region. Although patent data may most accurately reflect regional innovation activity by manufacturing structure, we omit it for two reasons: First, it would go beyond the scope of this study to assess the effects of particular industries. Second, data on some regions and sectors is of very low size – which we judge to bear too much discretionary effect. Therefore only total patent application figures are taken into account as mass indicator. As with the previously mentioned mass indicators, we expect patent applications to represent innovation activity and thus to promote FP collaboration.

General Eurostat macroeconomic statistics enhancing our data set comprise the regional totals of employees, population, GDP and GDP per capita in Euros and PPP (Eurostat 2003). These indicators may well reflect mass effects, but the primary purpose for their inclusion was to construct ratios with R&D data (see p. 110). Moreover the regional share of persons speaking English as a native or foreign language was compiled from European Commission (2001a). While language differences were already interpreted as a distance indicator (p. 101), proficiency particularly in English may represent additive effects on transaction/collaboration. Button et al. (1993) mention lack of English language skills as one of the major obstacles to international research collaboration. Therefore an elevated share of English speakers is expected to promote FP interaction, if there is a significant relationship.

All of the Eurostat mass indicators are unimodally distributed, but each of the resulting data vectors is subject to skewness and kurtosis not resembling normal distribution. In contrast,

[79] Furthermore, Eurostat (2003) provides the „Private Non-Profit“ (PNP) division, which constitutes less than 2% of research expenditure in EU-15 NUTS-1 regions (The only notable exception being mainland Portugal PT1 with 14%). Since data on PNP is infrequent and in case of R&D expenditure even not existent for some regions, it was excluded from our data set. Consequently, further reference to “total” R&D expenditure represents the sum of expenditure by the business, government and education sector.

standard tests do not reject the hypothesis of normal distribution for their log-transformed counterparts. To a lesser extent, this also refers to GDP/capita data. Furthermore, the mentioned mass indicators are highly collinear – i.e. (apart from GDP/capita) the absolute values of the respective correlation matrix elements are seldom below 0.7, and are even beyond 0.9 within the HRST, expenditure or personnel domains. This causes problems to the application of least squares methods, since collinear input factors may bias variance estimators.

Amadeus firm size data stems from queries at the Amadeus European business register (Amadeus 2003), which provides total entry numbers grouped according to region and firm size. The latter dimension comprises data on firms with less than 200 employees, with 200 to 500, with 501 to 1000 and with more than 1000 employees, totalling more than 5 million enterprises.[80] Furthermore Amadeus provides the same structure only counting firms categorised among the Amadeus-defined "European Top 1,500,000". Although we do not know exactly the criteria for inclusion into this selective class, we reckon that this category encompasses nearly all of prospective FP project participant firms. Moreover Amadeus coverage may differ by European country – a problem expected to be less pronounced with figures for Europe's most important firms. On p. 115 we will assess the relevance of both indicator groups. Apart from collinearity, two familiar problems arise with the appropriateness of the resulting data vectors for least squares analysis: Normal distribution is rejected for all Amadeus data and figures are distributed even less smoothly than Eurostat indicators. Due to the latter property the normality assumption for the log-transformed vectors is less clear-cut than for Eurostat data.

Firm numbers are certainly expected to exert positive influence on cooperation. Furthermore, the literature review in section 2.1 points out that propensity for research collaboration rises with firm size. Thus the positive relationship may even be more pronounced for the number of large firms, respectively their regional share. In addition, it is stated that differences in firm size may hinder collaboration. The firm size structure distance indicator on p. 103 tries to capture this effect.

5.3.3 Ratios – Efficiency Indicators

The hypothetical masses behind the FP gravity model are expected to depend of absolute numbers, as for instance GDP figures. Moreover, a node's propensity to collaborate may as well be influenced by certain efficiency indicators (e.g. GDP per employed person). The latter result from combinations of the absolute Eurostat and Amadeus mass indicators. In most

[80] The provided total figures include enterprises not attributable to a firm size class.

cases, these combinations consist of ratios, i.e. are multiplicative in original figures and additive in their log-transformed counterparts. Some of those efficiency measures are constructed according to suggestions by the literature and some are due to a-priori economic reasoning – But a considerable part stems from explorative least squares analysis, which tests various combinations of possibly explanatory factors for significance. If a set of two indicators exhibits an extraordinary linkage of regression coefficients, it may hint at a dependence of a single linear factor combination rather than that of both indicators at a time. Such a factor combination may represent an economic relationship not taken into account before. For the sake of readability, we present those indicators and the underlying economic reasoning among the "a priori" ratios listed below.

R&D ratios: Apart from the absolute size of R&D activity in a region, its relevance to the regional environment might matter. Relating sector-specific *R&D expenditure to GDP* depicts concentration effects and may as well reflect the quality of a region's R&D in business, governmental and educational research. The same reasoning goes for *R&D staff in relation to employment* numbers. It is unclear, however, which of and how those indicators may affect FP cooperation. From p. 77 we know that academic and public institutions are over-proportionally represented in FP projects in relation to their share of European R&D expenditure and staff. On top of absolute size, these R&D ratios might indicate the orientation of the regional R&D activities to the private or to the public sector, and thus depict the push factors to cooperation. If the private sector is of particular importance, public institutions are able to find many rewarding possibilities for intra-regional cooperation, whereas in less prosperous regions they may be more oriented towards academic excellence and international partners. The *share of a sector in total R&D expenditure*, or of its R&D personnel compared to the sum of researchers exhibit the same reasoning more directly without relating it to GDP or employment numbers. *Direct ratios between R&D expenditure or personnel in one sector to that of the other* capture the same effect in a different numerical representation.

Research expenditure per researcher is one of the first ratios that come to mind, although considerations on its potential effect on FP collaboration are less clear-cut. It indicates the resources available per individual – the more resources are directed to researchers (in PPP), the more excellent may be their output of research, and thus the more attractive the region is for international collaboration. However, the more resources are dispensed on an individual, the less there may be the need for additional funding (through the FP) and sharing of resources. In addition, we may identify differing relationships for the business, government and education sectors. Since many more contradicting thoughts on these ratios' impact come to mind, further reasoning will be postponed to the interpretation of least squares analysis results.

Patents per researcher or per PPP of research expenditure reflect an already tangible innovation output of a regional R&D sector. Since patents are primarily the domain of applied and private-financed research, it seems most sensible to relate patents to R&D resources in the business sector. This ratio may reflect the business sectors efficiency, but also its positioning of applied versus generic research. From the latter thought, we would rather expect a negative relationship between patent per resource ratios and collaboration intensity, due to the FP focus on pre-competitive research.

The FP decision process is inclined to ensure adequate representation by all EU member states, enhanced by cohesion motivations concerning weaker countries. Moreover, push factors resulting from the lack of alternatives are certainly more pronounced for countries with less dynamic innovation activity. This would imply that “mass” deficiencies in less developed countries would at least partially be compensated for by political considerations. However, these considerations affect mainly the national level, while regions dispose of less political clout to ensure their “adequate” share of projects. Therefore, deficits in regional R&D might be less compensated for than deficits in national R&D.[81] This would translate into “over-proportional” assignment of projects to weaker EU countries, while at the intra-national level, projects would be allocated to the strongest regions. Figure 7, for instance, shows that national economic cores dispose of the most elevated implied masses at the national level, while the international differences between those economic cores are less pronounced.[82] In order to depict this potential effect, we divide regional expenditure/staff by the respective country total, for all of the three research sectors.

General macroeconomic and business ratios: If scale effects matter for FP collaboration intensity, regional *GDP per worker* may be considered as a catchall measure for the development status of the regional economy and thus for the efficiency of the entire regional economic system. Broken down on the individual level, it could be interpreted as resources available per individual. Whether this translates into positive or negative impact on collaboration remains indeterminate.

The literature review highlighted the relevance of firm size structures for collaboration. On top of the total numbers of large firms etc., a linkage between their importance to the regional economy and the R&D sector may exist. The average number of *employees per firm* (i.e. the total number of employees by the total number of Amadeus entries) is one possibility to

[81] In order to assess political effects at the national level, we introduce country dummies.

[82] Dividing implied masses by regional R&D personnel or R&D expenditure (not displayed) yields particularly high numbers for Spanish, Greek, Irish and Portuguese regions. The relevance of economic cores is however less obvious.

measure average firm size per region. The *share of firms with less than 500 employees classified among the Amadeus Top 1,500,00* seeks to indicate the innovativeness of regional small and medium enterprises (SME). Similarly, dividing Amadeus Top firms by the total of Amadeus entries should represent the private sector's general innovation resources. *Research intensity by medium and large enterprises* is measured by relating business expenditure to the number of top firms with more than 200 workers.

All of the firm-size-related indicators above are expected to affect FP collaboration positively (if they do have an impact). The same is true for the number of *top firms with more than 1000 employees related to the total number of firms with less than 200* employees, created to distinguish regions characterised by large corporations versus SME-dominated structures. We expect it to exert positive influence on cooperation figures, since large firms are more inclined to FP collaborations, whereas small innovative firms may find enough cooperation opportunities at the regional and national level.

The ratios presented are largely unimodal, but their histograms frequently appear skewed to the left, even if they range between 0 and 1. Therefore log-transformation is as well an attractive option for those numbers. In this case, however the transformation consists of adding 0.01 rather than 1 before taking logs, since adding 1 would bias the respective logarithmic distributions even more. For further details, please refer to the descriptive statistics in Table A.6 in the appendix.

5.3.4 Relatedness among Mass Indicators

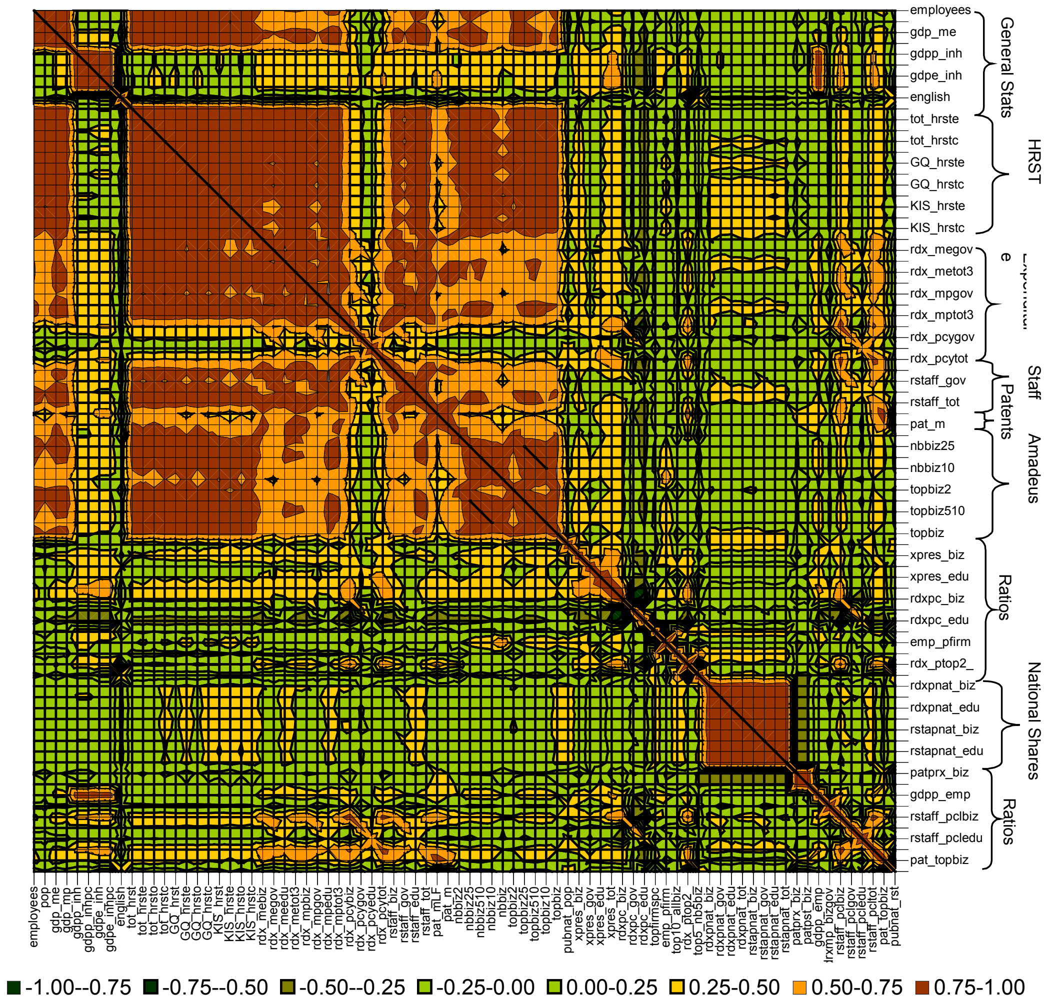
Since the mass indicators presented constitute the explanatory variables for least squares analysis, the degree of their collinearity is of tremendous importance. Two variables are regarded as collinear, if they are highly related: The more two vectors are linearly dependent, the more the two are correlated.[83] An inclusion of collinear regressors in least squares analysis results into biased estimates for the variance of coefficients, and thus leads to wrong conclusions about the significance of factors (Greene 2003, p. 260).

Figure 8 portrays the correlation matrix between the mass indicators used in this study. As expected, substantial correlations can be identified particularly among indicators in absolute numbers (the upper left part of the graph). Moreover, these correlations are especially pronounced among indicators of one group (e.g. HRST).

---

[83] Equation (26) demonstrates how to measure linear dependence between two vectors x and y. If applied on centralised vectors (i.e. vectors $\mathbf{x}-\overline{\mathbf{x}}$ and $\mathbf{y}-\overline{\mathbf{y}}$), the cosine of the angle between the two vectors corresponds to their squared correlation coefficient by definition.

Figure 8: Correlation matrix among 80 mass indicators, graphic representation[84]



Ratios exhibit less correlation with other indicators, apart from correlations within groups of ratios (e.g. expenditure per researcher in the government sector versus the business sector). The overall multi-collinearity of the data set vectors is expressed by the fact that the largest eigenvalue of the correlation matrix is 36.57 – compared to the total number of factors (80), this implies that one eigenvector already explains 46% of correlation matrix variance. Concerning the subset of ratios (lower, right part of the matrix), still 31% of the matrix is described by the "first" eigenvector.[85]

---

[84] Indicators are sorted as follows: 1 - upper/left part: variables directly supplied by data contributors (employees to *PUBNAT_POP*); 2 – lower/right part: ratios compiled thereof (*XPRES_BIZ* to *PUBNAT_RST*). For further details, please refer to Table A.6 in the appendix.

[85] Regarding both the entire matrix and the "ratio subset", the 12 most important eigenvectors explain roughly 97% of variance.

The present multi-collinearity among mass indicators constitutes a certain danger to the estimation methods performed later on. In order to reduce this risk, we will examine similar variables one by one – for instance to determine the "better indicator" by separately analysing GDP in PPP and in Euros.

5.3.5 Appropriateness of Mass Indicators

In order to assess the explicative of power of mass indicators prior to final estimation steps, we compile their correlations with several derivatives of the FP cooperation matrix – as depicted in Table A.7 in the appendix. The first eigenvector, implied masses and row sum series represent inter-related measures of a large part of mass effects and bear substantial correlation with the absolute figures presented above – virtually all correlations with those range between 0.6 and 0.8. Additional "efficiency" effects may be found in the remaining eigenvectors, which seem to bear resemblance to several ratio indicators. Most notably, the third eigenvector exhibits correlation coefficients of more than 0.6 with patent-related variables. In general, the ratio indicators' frequent correlation coefficients of more than 0.3 hint at explicative power of ratios on top of "normal" mass indicators.

Moreover, Table A.7 allows reducing the range of variables to be examined in the further course of this study: Within a group of similar variables, only a single series may be included into LS regressions for multi-collinearity reasons. For instance, the total number of Amadeus top firms (*TOPBIZ*) seems to reflect mass distribution more effectively than the total number of Amadeus firms (*NBBIZ*). R&D expenditure expressed in PPP exhibits greater correlation coefficients than in ECU. Human Resources (HRST) for the service sector ("GQ") and for knowledge-intensive services ("KIS") perform better than overall HRST. Several more indicators might be identified as superior to others, although differences are less clearly to determine. Several indicators (e.g. *TOPFIRMSPC*) exhibit minor correlation coefficients, thus they may justifiably be omitted in further analysis.

A narrower data set is of particular importance to the methods presented in section 6.1, while the whole range of data may be tested in other cases.

5.3.6 Fixed Effects – Dummy Variables

In addition to collaboration patterns due to empirical mass indicators, there might be substantial "fixed effects" associated with peculiar regions. When modelling implied masses by the mass indicators presented, the resulting error vector may provide insights into region-specific effects. Furthermore, there may be effects attributed to countries rather than regions. Greece, for instance, is often mentioned to be over-proportionally involved into FP projects – particularly if adjusted for its under-proportional R&D resources. In general, smaller countries may be more inclined to FP cooperation than their larger counterparts. Looking at implied masses, the Eastern German *Länder* seem to have been particularly infrequently involved into the FP4.

These specific effects are captured by dummy variables, adding as a constant if the dependent observation is thought to belong to such a peculiar grouping. For the example of Greece, this widely used method would lead to n x 1 data vector whose elements take the value 1 if the respective observation is a Greek region, and 0 otherwise. Subsequently, this vector would be added to a model of implied masses in order to examine Greece-specific effects. The same methodology is applied to other regional groupings containing more than a few nodes: Namely the dummy factors for Germany, Spain, France, Italy, the UK, Greece, the Netherlands and the Nordic countries. Furthermore, we use dummies referring to the 10 small member states, to the 5 Eastern German *Länder* (ex-Berlin), and to capital regions.

With analysis of stacked data (as described in (27)), the dummy variable referring to Greece may be imagined to derive from an inter-node matrix: The rows and columns exhibiting a node pair with at least one Greek region value 1, while the remaining elements are 0. This matrix results in a stacked dummy vector, whose elements are 1 if a Greek region is involved in the respective observation. With stacked data, the study of regional rather than country fixed effects requires dummy factors as well, since n observations are attributable to each single region.

6 A GRAVITY MODEL OF RESEARCH COLLABORATION

During the following section, we will assess the impact and explanatory power of exogenous mass and distance indicators regarding FP collaboration. For this purpose, we resort to two approaches: First, we take exogenous distance data to estimate the implied distance matrix, and model implied masses with mass indicators. Second, we try to model the gravity formula (as in (3)) directly by modelling observed collaboration with exogenous variables. A concluding section will compare the results of both approaches and try to interpret them in economic terms.

Before, this section lays out several econometric considerations concerning the modelling process: We rely on Least Squares (LS) methods to carry out the model structure finding process. However some constraints are attached to LS methods, chiefly concerning the properties of model error terms. We will introduce several adjustments in order to ensure at least broad compliance to those constraints. Since this study has exploratory character, its main technical focus lies on the design of the prospective variable evaluation procedure. In particular, we require estimation outcomes to exhibit stable relationships regardless of geographic region. The latter sections on implied distance and masses modelling, as well as direct modelling will put these considerations into practice.

6.1 Modelling Procedures and Requirements

Both the direct and the indirect approach to modelling regional collaboration numbers are carried out with the Least Squares estimation method.

6.1.1 Data Stacking with Symmetric Matrices

Section 5 leaves us with several symmetric n x 1 matrices and n x 1 vectors for use as model inputs. First, dependent factors (i.e. the data we aim to model) consist of the symmetric total collaboration matrix **T**, and its derivatives: the implied distance matrix **D*** and the implied masses vector **m***. Second, we introduced explanatory variables ${}_i\mathbf{x}$, ${}_i\mathbf{X}$: Those classified under the "distance" category appear as symmetric n x n matrices ${}_i\mathbf{X}$, while "mass"-related indicators appear as n x 1 vectors ${}_i\mathbf{x}$. Conforming to (2), respectively (4), we use the gravity model to explain the first matrices by the latter – as in (28).

(28) $\mathbf{T} = \mathrm{f}(\theta, \mathbf{mm}', \mathbf{D}) \qquad \mathbf{m} = \mathrm{f}({}_1\mathbf{x}, {}_2\mathbf{x}, ...) \quad \mathbf{mm}' = \mathrm{f}({}_1\mathbf{x}_1\mathbf{x}', {}_2\mathbf{x}_2\mathbf{x}', ...) \;\; \mathbf{D} = \mathrm{f}({}_1\mathbf{X}, {}_2\mathbf{X}, ...)$

Either we interpret **T** to be constituted of implied masses and implied distances, and try to model those implied factors. Or we integrate the formulas from (28) into (29) and estimate the collaboration matrix directly with exogenous distance and mass indicators.

(29) $\mathbf{T} = \mathrm{f}(k, {}_1\mathbf{x}_1\mathbf{x}', {}_2\mathbf{x}_2\mathbf{x}', ..., {}_1\mathbf{X}, {}_2\mathbf{X})$

Both the latter and the former approach may be modelled with least squares methods, implying a linear model structure as presented below (for direct estimation of T).

$$T_{i,j} = c + \alpha_1 \mathrm{f}({}_1x_i\,{}_1x_j) + \alpha_2 \mathrm{f}({}_2x_i\,{}_2x_j) ... + \beta_1\, \mathrm{f}({}_1d_{i,j}) + \beta_2\, \mathrm{f}({}_2d_{i,j})$$

Since this structure would hold for all i,j analysed, we are able to rewrite it for entire matrices:

(30) $\mathbf{T} = c\mathbf{1} + \alpha_1\, \mathrm{f}({}_1\mathbf{x}_1\mathbf{x}') + \alpha_2\, \mathrm{f}({}_2\mathbf{x}_2\mathbf{x}') + ... + \beta_1\, \mathrm{f}({}_1\mathbf{X}) + \beta_2\, \mathrm{f}({}_2\mathbf{X}) + ...$

The linear relationship presented in (3) represents n*n equations reflecting the same linear relationship for i,j. Since the entire set of matrices analysed is symmetric, the upper triangular part of these matrices is in fact redundant. This implies that the n*n equations reduce to n * (n + 1) /2, i.e. the main diagonals and lower triangular matrices in (30).

The resulting linear structure may easily be estimated with standard statistical methods, for instance by performing least squares on the explanatory variables functions f(x). These methods require variables to be laid out in vectors rather than in the matrix form they appear in with modelling total cooperation **T** and implied distances **D***. Therefore we "stack" the matrix data according to (27):

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1n} \\ D_{12} & D_{22} & \cdots & D_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ D_{1n} & D_{2n} & \cdots & D_{nn} \end{pmatrix} \Rightarrow \overline{D} = \begin{pmatrix} D_{11} \\ \vdots \\ D_{1n} \\ D_{22} \\ \vdots \\ D_{2n} \\ \vdots \\ D_{nn} \end{pmatrix}$$

Stacking enables the operation of standard methods while retaining the form of the n * (n + 1) /2 equations. Log-transformation of variables was introduced to convert the multiplicative gravity structure in (2) to the linear one in (4). The functions f in the representations above reflect the necessary log-linearisation of matrix elements – moreover, the may include further operation on data in order to render its properties well behaved (and in particular unskewed).

6.1.2 Problems with Least Squares Estimation

In order to estimate the gravity model structure by the means of Ordinary Least Squares (OLS), the model factors have to conform to the Gauss-Markov conditions (Greene 2003, p. 56). According to Gauss-Markov requirements, an OLS estimator is the most efficient (i.e. the one with minimal coefficient variance) only if:

1) the rank of the design matrix X (containing the regressor data in its columns) is equal to the number of explanatory factors
2) the explanatory factor data is fixed rather than stochastic (i.e. there is no error term in explanatory data)
3) the error terms' expected values are zero, their variance is even for all errors (i.e. they are identically distributed) and they are not correlated (i.e. independently distributed).

(31)
$$\begin{array}{l} 1)\ \mathrm{r}(\mathbf{X}) = k \\ 2)\ X_{i,j}\quad \mathit{fixed} \\ \left.\begin{array}{l} 3)\ \mathrm{E}(\mathbf{u}) = 0 \\ 4)\ \mathrm{Cov}(u_i, u_j) = 0 \\ 5)\ \mathrm{Var}(\mathbf{u}) = \sigma^2 \mathbf{I} \end{array}\right\} \Rightarrow u \sim \mathrm{IID}(0, \sigma^2) \end{array}$$

$\mathbf{X}$...matrix of regressors (in columns)
k...number of columns of $\mathbf{X}$, $X_{i,j}$...element of $\mathbf{X}$
$\mathbf{u}$...vector of error terms, u_i...element of u
σ^2...variance of u
IID ..." Independently Identically Distributed"

The first Gauss-Markov condition implies that none of the factors used are linearly dependent, which would render the inversion of **X'X**, and thus OLS estimation, infeasible. Preferably the columns of **X** would be orthogonal, i.e. they should exhibit a correlation of zero among them. With economic data this is hardly the case: frequently considerable linear relationships between regressor variables exist, which would imply multi-collinear variables. Multi-collinearity leads to biased estimators for coefficients and coefficient variance. In order to reduce this risk, we will aim at keeping the correlation coefficients between the regressors included as low as possible.

The second requirement, namely to regress on fixed rather than stochastic variables, renders the inclusion of most economic data debatable. However, most economic studies ignore this problem, since explanatory data is assumed to reflect the indicators it is describing with sufficient accuracy.

The requirement of independently and identically distributed data is somewhat more demanding with respect to the data used in this study. Although the mean of the observed OLS residuals is equal to zero by definition this may not necessarily imply that the expected value of the error term to be estimated is zero as well. In particular, the error term's expected value may differ between subsets of the matrices analysed. For instance, node-specific factors unaccounted for may force the expected values of errors to differ with respect to nodes. I.e. a certain model structure may thoroughly overstate collaboration figures for node i while understating those for node j. [86] These fixed effects may easily be corrected for by adding n=68 node-specific regional dummy variables – equivalently to introducing node-specific constants.[87]

If the variance matrix Var(**u**) does not conform to the structure $\sigma^2 \mathrm{I}$, its estimators may be biased, which in turn may lead to wrong conclusions regarding the significance of parameters. One source for such misbehaviour is heteroskedasticity where variance σ^2 is not uniform for all elements of **u** but differing σ^2 for each observation i,j. This would imply a variance matrix Var(u) = $\mathrm{Diag}(\sigma^2_{1,1},\sigma^2_{2,1},\ldots\sigma^2_{2,2},\sigma^2_{2,3},\ldots\sigma^2_{i,j},\ldots)$.[88] In OLS equation structures, the existence of heteroskedasticity can be tested for by standard methods, e.g. the White test (Greene 2003, p. 222). In case heteroskedasticity exists but its structure is unknown, it is

---

[86] This would imply $E(u_{i,k}) \neq E(u_{j,k}) \neq 0$ for $i \neq j$ and $\forall$ k.

[87] Introducing n dummy variables refers to the estimation of total collaboration **T** and implied distances **D***.

[88] Keep in mind that the variance matrix of stacked data is of dimension n (n+1)/2 x n (n+1)/2, i.e. 2346x2346.

recommendable to apply White's heteroskedasticity-consistent variance estimator to correct biased standard errors.

In case the variance structure is known, however, Weighted Least Squares (WLS) may be used to correct for heteroskedasticity. If a data series proportional to the non-observable variance series ("weighting series") can be identified, the condition of homoskedastic errors may be achieved by dividing the dependent variable and the regressor variables through the weighting series. The such-obtained coefficients then are applied to describe the model structure with respect to unweighted data. However, Greene (2003, p. 219) points out that the weighting series has to be at least strongly related to one of the variables used in estimation. If there is no relation between weighting series and estimation data, OLS computations "will not be misleading" (Ibid.).

Even if the underlying weighting series may be known, its precise form for inclusion may not be correctly specified. Greene (2003, p. 226) concludes that incorrect specification may lead to artificial constants on both sides of the estimated equation and thus may overstate the R^2 obtained by WLS. Therefore he recommends dismissing WLS structures with R^2 larger than the R^2 obtained by OLS estimation.

Bearing these points in mind, we consider some variance structures suspected within the presented data set: As with error expected values, differences in variance may be imagined to depend on node-specific effects, i.e. a node i may be associated with larger variance than node j, while the $\sigma^2_{i,j}$ of the collaboration/distance between both is a product of both node-specific variance terms $\sigma^2_{i,j}= \sigma_{i,i}\ \sigma_{j,j}$.[89]

It is not exactly determined how this group-wise heteroskedasticity is structured, i.e. which nodes exhibit larger variance than others. A first approach would be to estimate the model in "regular" structure by the means of OLS and relate the squared residuals to the variables in the data set. We have put this approach into practice, but since it did not yield significant quality enhancement, it will not be discussed for the sake of briefness.

A-priori, we might assume a relationship between the dependent variable (namely either inter-nodal FP collaboration links or implied masses and implied distances) and variance: Several nodes exhibit very large collaboration numbers while the cooperation intensity of others is extremely low. In absolute terms, collaboration intensity may increase with mass – however this effect is likely to be balanced by the log-transformation of mass variables. In

[89] The variance assumption for bilateral observations concerns the estimation of cooperation **T** or implied distances **D*.** Implied masses **M***, in contrast, do not exhibit bilateral elements and would therefore only be subject to independent variances $\sigma_i^{\ 2}$.

contrast, variance relative to mass size might be inversely related to masses. The less a node's number of total collaboration links, the more the discrete nature of the dependent variables surface, and the more collaboration intensity is distributed in an unbalanced manner.[90] Out of total collaboration figures a weighting series is first constructed as a symmetric matrix: Row sums (i.e. total collaboration per node) form the main diagonal of this matrix, i.e. the $\sigma^2_{i,i}$. The remaining matrix elements $\sigma^2_{i,j}= \sigma_{i,i}\ \sigma_{j,j}$ derive from multiplying the square roots of the main diagonal elements i and j.

Instead of mass size, the core-periphery positioning of nodes may even more influence variance in collaboration error terms. The farther a node is apart from the centre of regional collaborative networking, the less it is relatively involved with more distant nodes, the more discrete and spontaneous and thus the more its collaboration with those nodes might vary in intensity. The row sums of implied distances would be an indicator of first-choice to represent such multilateral core-periphery standings. However, the same argument applies to bilateral distances as well: Instead of distance row sums, the direct application of distance indicators may yield even better results.

The squared residuals of a regular OLS model may evenly be applied for WLS: Adding up the residual squares per node yields a proxy of node variance. If these variance indicators form the main diagonal of a symmetric $n \times n$ matrix, the remaining elements may be computed out of the main diagonal as above.

However it is not clear whether node-specific variance actually exists and whether the a-priori mentioned weighting series might accurately reflect this variance. Therefore we will continue steps with displaying both the performance of OLS and WLS structures.

6.1.3 Consistency of the Estimated Structure

While may the approach sketched above may lead to apparently good results, it is questionable whether good performance indicators only stem from extraordinary fit in one part of the sample. Comparing a model structure with data hitherto not used for estimation might reveal this „over-fitting“. With available data, we unfortunately are not able to test the stability of estimated model structures over time. Nevertheless, with our extensive data set we are at least able to test stability over the geographic dimension. I.e. we assume a model to be stable if it does not change its structure with respect to random node sampling.

For this purpose we partition the sample of dependent and explanatory factors into to independent sets of equal size. Then the model structure in question is fitted to both sets,

[90] A comparison between row sums (or implied masses) and row variances of the total collaboration matrix shows this relation to be near to non-existence.

and the resulting two estimations' outcomes are checked for consistency by applying the "Chow forecast test".

Example (32) explains how an "independent" data set is chosen. Out of the n=68 nodes with their n (n+1) /2 bilateral combinations, two sets of m nodes each are randomly drawn. This results into a two data sets, each with m nodes and m (m+1) /2 cooperation links among them.[91] Cross-set links between nodes of both data sets are not included. Therefore each drawn set contains its own closed collaborative sub-matrix. Example (32) describes an n=4 node matrix to be divided into a set with nodes {4,1} and another with nodes {2,3}.

(32) $$\text{Nodes}\{1,2,3,4\}..\begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{21} & x_{22} & x_{32} & x_{42} \\ x_{31} & x_{32} & x_{33} & x_{43} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix} \Rightarrow \text{Nodes}\{4,1\}...\begin{pmatrix} x_{44} & x_{41} \\ x_{41} & x_{11} \end{pmatrix} \ \text{Nodes}\{2,3\}...\begin{pmatrix} x_{22} & x_{32} \\ x_{32} & x_{33} \end{pmatrix}$$

The Chow forecast test originally was developed for testing model stability over time, but may evenly be applied to groups clustered along other dimensions. According to Greene (2003, p. 353) we construct the Chow forecast F-statistic as follows:

(33) $$F = \frac{(\mathbf{\bar{u}'\bar{u}} - \mathbf{u'u}) / (m(m+1)/2)}{\mathbf{u'u} / ((m(m+1)/2) - k)}$$

While $\mathbf{\bar{u}'\bar{u}}$ represents the squared sum of residuals of the estimation over the two combined subsets, $\mathbf{u'u}$ is the residual sum of squares of the first subset. The statistic is adjusted for the number of observations by subset m (m+1) /2 as well as for the number of parameters k. Equivalently, the resulting statistic follows F-distribution with m (m+1) /2 -k numerator degrees of freedom and m (m+1) /2 denominator degrees of freedom. A large F-statistic diminishes the chance of error when rejecting hypothesis zero (that the model structure is equal over both subsets).[92]

Our purpose is therefore to find a model structure that reaches sufficiently low Chow forecast statistics over different random subsets. For each model structure in question, the procedure and Chow statistic computation is repeated several times, in order to determine stability, and thus consistency, of a model structure.

---

[91] With implied masses **M*** estimation, data vectors contain only n=68 observations, thus the subsets drawn contain m nodes and thus m observations. Inter-node links are not relevant in this case.

[92] With implied masses **M***, the expression m (m+1) /2 reduces to m.

It has to be noted that this approach to model structure is not feasible with NUTS-1 dummy variables. Since subsets are separated according to nodes, a regional dummy (say AT1) would only be included into one subset, while its value would be zero over the entire range of the other subset. With regionally agglomerated variables (say an “AT” dummy for AT1, AT2 and AT3), compiling subsets with this dummy being 0 and 1 in both subsets would be possible. However, with only a few regions included in such an agglomerated dummy, the range of possible randomly drawn subsets would be greatly reduced. Therefore we only include agglomerated dummies with at least a sufficiently large number of corresponding regions into model structures assessed by the described procedure.

6.1.4 Evaluation of Explanatory Variables

Based on the explanatory data set defined in sections 5.2 and 5.3, we aim at identifying those factors providing best explicative power in estimation and stability according to the Chow forecast test. In a first step, irrelevant variables may be filtered out by analysing simple correlations with the dependent variable and its derivatives. Still, this leaves a relatively large (and collinear) set of variables to be tested in further analysis.

On top of correlations, we are interested in a factor’s performance in inter-relationship with other factors – think, for instance, of efficiency variables, whose impact is rather felt in combination with “plain” mass indicators. Therefore the second step of factor evaluation consists of testing every possible double combination of the variables having “survived” correlation-based filtering. Double after double, those combinations are added to a basic model structure (containing constants and dummy variables) and are assessed for model performance and stability.

With each double, the method described in section 6.1.3 produces Chow forecast statistics for numerous random subsets. Mainly according to those stability indicators, but also on the basis of t-statistics and fitting errors, variables are selected for further examination. This final step basically consists of manually evaluating regression performance and Chow statistics for several combinations of the remaining factors. Compare Table 8 for a general blueprint of the evaluation process, and see Procedure 1 to Procedure 9 in the appendix for a technical description.

Table 8: Outline of eligible variable evaluation procedure

1	Consider each possible variable combination (= nbvar (nbvar-1) /2 where nbvar is number of variables)
2	For each variable double:
3	Perform a regression of the dependent on the variable double and fixed factors (e.g. constant, dummies, etc.)
4	For a predefined number of times:
5	Divide dependent and exogenous data into two randomly constituted, independent sub-samples, and estimate the same model structure on both sub-samples
6	Perform Chow forecast test on the models obtained by both subsets
7	Store forecast statistic, and binary indicator whether individual coefficients are significantly different from zero
8	Return to step 4
9	Store average Chow forecast statistics and t-statistics in table “fstats”
10	Return to step 2
11	Present total share of insignificant Chow stats and significant t-statistics for each variable pairing in table “mc_results”

6.1.5 Scaling Final Output

Both direct estimation of **T** and modelling **T** through the estimation of **D*** and **M*** will yield modelled bilateral cooperation figures $\hat{\mathbf{T}}$. The absolute performance in estimating **T** is easily measured by regarding the fit’s residuals $\hat{\mathbf{T}} - \mathbf{T}$ and applying standard statistics such as adjusted R^2.

Apart from absolute model residuals, we are interested in whether a model provides a reasonable explanation of the inter-regional cooperation structure in relative terms. For this purpose, we scale $\hat{\mathbf{T}}$ by the its observed regional (row) sums. The procedure adjusts for errors in the modelling of absolute cooperation numbers, i.e. the absolute mass effect, and focuses instead on the performance of modelling bilateral connections.

The method is oriented at the “doubly constrained gravity model” approach outlined by Haynes/Fotheringham (1988, p. 25), which aims at scaling the nodes’ estimated masses to attain an interaction matrix exhibiting pre-defined sums over rows and columns.

Equivalently, we rescale the estimated symmetric transaction matrix $\hat{\mathbf{T}}$ to compile the scaled model output $\hat{\mathbf{T}}^*$. This scaled matrix exhibits the same vertical/horizontal sums of columns/rows c_i than does the observed matrix **T**.

(34) $$c_i = \sum_i T_{ij} = \sum_j T_{ij} = \sum_i \hat{T}^*_{ij} = \sum_j \hat{T}^*_{ij} \qquad \forall\, i,j$$

In order to comply with this constraint, the elements of the estimated matrix $\hat{\mathbf{T}}$ are iteratively scaled according to procedure (35). Each iteration consists of four steps: The first iteration’s (t=0) first step compiles the column sums of the initial matrix $\hat{\mathbf{T}}^*_{\mathbf{0}}$. Dividing the constraint vector’s elements $c_{0,i}$ by the obtained sums $b_{0,i}$ results in the scaling vector $\mathbf{a_0}$. The elements of this scaling vector are imputed into the main diagonal of diagonal matrix $\mathbf{A_0}$.[93] This scaling matrix times the transposed initial matrix $\hat{\mathbf{T}}_{\mathbf{0}}'^{*}$ yields the scaled matrix $\hat{\mathbf{T}}_{\mathbf{1}}^{*}$ at iteration step t=1. This matrix already exhibits column sums equal to $\mathbf{c}$, but its row sums still do not fulfil this condition. Therefore iteration step 1 calculates the row sums $\mathbf{b_1}$. The procedure continues until the elements of vector $\mathbf{a_t}$ come sufficiently close to one (or until $\mathbf{b_t}$ resembles $\mathbf{c}$, respectively). Note that the transposing of $\hat{\mathbf{T}}_{t}'^{*}$ in the first and last step of each iteration ensures that rows and columns are scaled alternately.

$$\hat{\mathbf{T}}_{\mathbf{0}}^{*} = \hat{\mathbf{T}}$$

$$\textbf{(35)}\quad \mathbf{b_0} = \hat{\mathbf{T}}_{\mathbf{0}}'^{*}\vec{\mathbf{1}} \qquad \mathbf{a_o} = \left\{ c_{0,i} \middle/ b_{0,i} \right\} \qquad \mathbf{A_0} = \mathrm{Diag}(\mathbf{a_0}) \qquad \hat{\mathbf{T}}_{1}^{*} = \mathbf{A}\hat{\mathbf{T}}_{\mathbf{0}}'^{*}$$

$$\mathbf{b_t} = \hat{\mathbf{T}}_{\mathbf{t}}'^{*}\vec{\mathbf{1}} \qquad \mathbf{a_t} = \left\{ c_{t,i} \middle/ b_{t,i} \right\} \qquad \mathbf{A_t} = \mathrm{Diag}(\mathbf{a_t}) \qquad \hat{\mathbf{T}}_{\mathbf{t+1}}^{*} = \mathbf{A}\hat{\mathbf{T}}_{\mathbf{t}}'^{*}$$

The final scaled output $\hat{\mathbf{T}}_{\mathbf{T}}^{*} = \hat{\mathbf{T}}^{*}$ at iteration t=T may equivalently be obtained by multiplying the initial matrix with the main diagonal's square roots of a scaling matrix **A*** - see (37). As depicted in (36), **A*** is the product of the $\mathbf{A_t}$ obtained during iteration.[94]

$$\textbf{(36)}\quad \mathbf{A}^{*} = \mathbf{A_0 A_1 \ldots A_t \ldots A_T}$$

$$\hat{\mathbf{T}}_{\mathbf{T}}^{*} = \left\{\sqrt{a_{ii}^{*}}\right\} \hat{\mathbf{T}}_{\mathbf{0}}^{*} \left\{\sqrt{a_{ii}^{*}}\right\}$$

$$\textbf{(37)}\quad \left\{\sqrt{a_{ii}^{*}}\right\} = \begin{pmatrix} \sqrt{a_{11}^{*}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{a_{nn}^{*}} \end{pmatrix} \qquad \left\{a_{ii}^{*}\right\} = \mathbf{A}^{*}$$

Standard statistic performance indicators may then be applied to the result of the scaling process and thus indicate the goodness of fit for internal matrix structure.

---

[93] For a given vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, the diagonalised matrix is $\mathrm{Diag}(\mathbf{x}) = \begin{pmatrix} x_{11} & 0 \\ 0 & x_{22} \end{pmatrix}$

[94] Since with increasing t $\mathbf{A_t}$ converges towards the unity matrix, **A*** converges towards a stable solution.

6.2 Modelling Implied Distance

The following two sub-sections will be dedicated to the modelling of implied masses and distances: First, we will focus on distances and assess its explicative value in combination with row sums of the collaboration matrix (as a proxy for masses). Then we will model implied masses independently and re-combine it with implied distances.

6.2.1 Modelling Implied Distance's Core-Periphery Element

Implied distances, a collaboration matrix derivative, were introduced in section 5.1.2 and their distribution was examined in particular with respect to regional nodes. In section 5.2, we discussed potentially relevant "exogenous" factors related to implied distance. Some of those exogenous distance matrices were directly obtained from primary data, whereas most of them result from the combination of several corresponding $n \times 1$ vectors, which would by themselves rather be classified as masses. Although the aim of our modelling approach is to separate between node-specific and bilateral effect, the line between those is somehow blurred: Keep in mind that implied distances derive directly from the collaboration matrix in order to extract purely bilateral effects, but nevertheless contain information about the core-periphery positioning of nodes. The latter factor can be compressed into an $n \times 1$ vector describing "distance to the centre". And this $n \times 1$ vector in turn might be shaped by variables classified as "mass" factors.

The dual aim of this study is to find a model structure both exhibiting goodness of fit and stability over regional sub-samples. Regional dummy variables are introduced to measure regional fixed-effects, adjusting for over- or under-estimated distances by a general model structure. With implied distances, those regional dummies' coefficients constitute an indicator for the core-periphery positioning. Since our aim is to *explain* as much as possible of the dependent variable, we try to find a structure enough general to reduce the impact of fixed effects to the minimum.

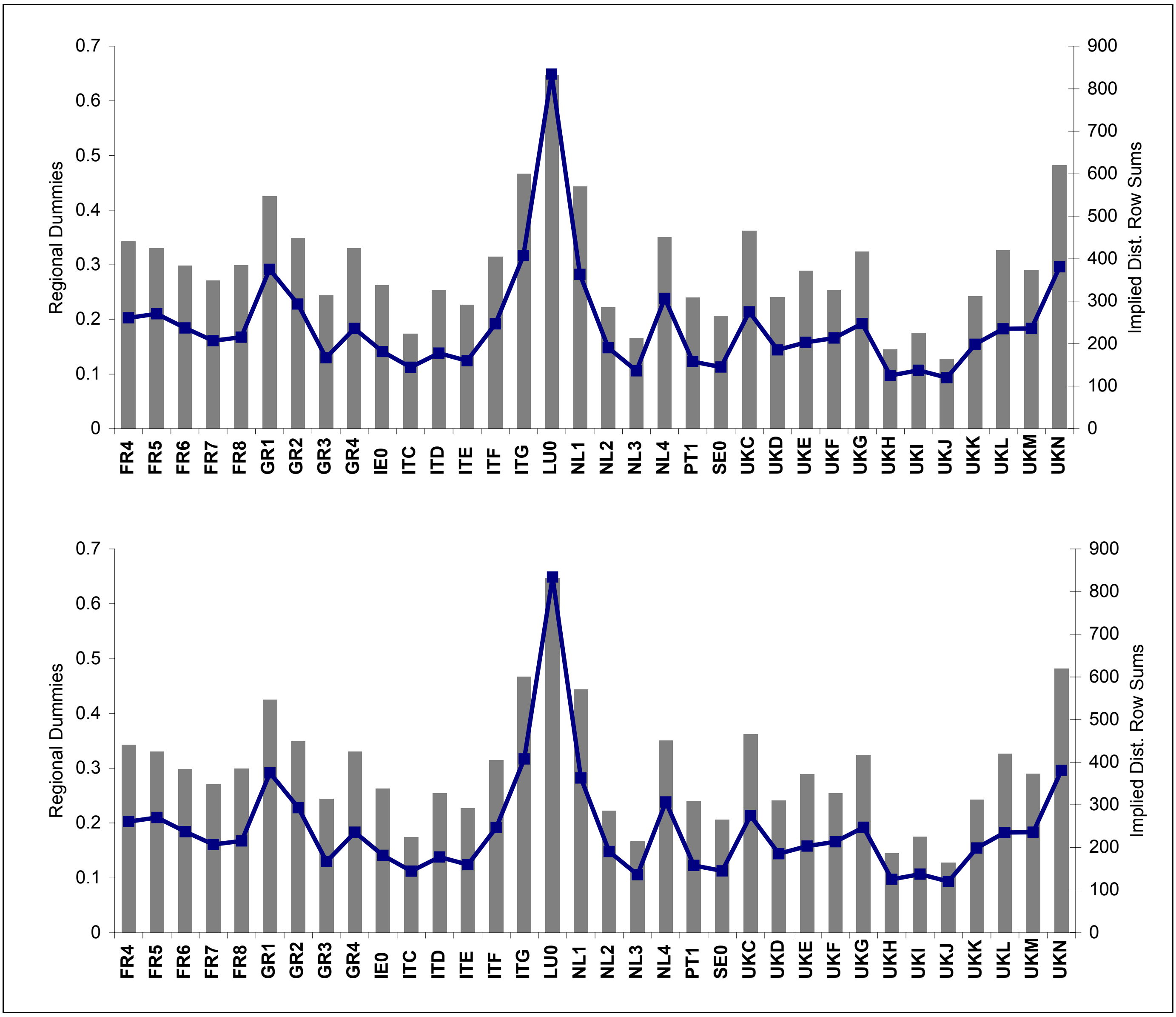
Having added the later goal to our modelling objectives, we start the estimation procedure (displayed in Table A.9, p. 191) with regressing implied distances on n=68 regional dummy variables in order to determine the importance of node-specific factors. The dependent variable, stacked implied distance, was doubly log-transformed in order to achieve linear values vaguely following normal distribution.[95]

---

[95] Double log-transformation results in the identifier *LL_IMPDIST*=log(log(*IMPDIST*+1)+1). In order to achieve skewness and kurtosis similar to normal distribution, this variable would in fact have been log-transformed several more times – we restricted ourselves to double transformation since the necessary addition of 1 prior to each transformation leads to data blurring.

The estimation output exhibits an adjusted R^2 of 0.64, revealing the importance of node-specific fixed effects in the implied distance matrix structure. Since we suspect those regional "constants" to reflect a node's core-periphery positioning, we relate the $n \times 1$ coefficient vector to the row sums of implied distances (doubly log-transformed). The comparison yields a remarkable correlation coefficient of 0.98, a relationship visualised by Figure 9.

Figure 9: Regional dummy coefficients in implied distance estimation [96]



As depicted in Figure 9, the dummy coefficients correspond to the regions' core-periphery positioning in the "implied distance map" (Figure 5, p. 93). The coefficients confirm that the

[96] Dummy coefficients stem from regressing a 2278x1 vector obtained by stacking the lower triangular of the implied distance matrix (excluding main diagonal elements) on 68 regional dummy variables.

regions regarded as the most peripheral within countries also exhibit the largest implied distances in FP collaboration (see concluding section 7 for further interpretation).

In order to assess the explanatory value of exogenous data, we envisage to apply the stability test on factor doubles sketched in section 6.1.4. As has been mentioned before, we would like to measure explicative value on top of fixed effects. However, the outlined factor assessment procedure is not feasible with regional dummies, since separating the data set into to unconnected samples would result in zero-vectors for half of each subset's dummies – rendering LS estimation impossible. In contrast, the fitted values of the fixed effects structure would add the same explicative value to a model structure but would allow for stability testing.[97] There are caveats though: The fixed effects structure is not yet adjusted for bilateral distance indicators, and embodies not only node-specific effects, but as well information on bilateral connections. Distance matrix estimation solely based on implied distance row sums, in contrast, explains nearly as much as the fixed effects model, but its fitted value matrix can be broken down into a single n x 1 vector (row sums). We prefer modelling this rather general periphery measure by several other n x 1 data vectors.

In order to determine the structure of the row sum model, we first looked at correlations between the (log-transformed) row sums and similarly constructed row sums of geographical distance, GDP-weighted distance, as well as the total set of explanatory "mass" variables, such as regional research staff. Since the intra-country core-periphery settings seem to matter, we introduced a further variable (*dist2core*) containing the intra-country geographical distance to the NUTS-1 region with the largest GDP. We then tried out the most likely combinations iteratively, optimising for parameter significance and stability.[98]

The resulting model structure contains four data vectors: The most important is (log-) total research staff per region, an undisputable "mass" indicator, which exerts negative effects on the periphery indicator.[99] The positive parameter for "distance to core" confirms our suspicion that intra-country periphery plays an important role to implied distances (in particular, more important than distance to the European core regions).

---

[97] Since our aim is to estimate the entire set of implied distances, we do not include its row sums among the regressors, since this indicator is a direct derivative of implied distances.

[98] Stability testing in this case means randomly sorting nodes/observations in the data set and performing a Chow forecast test for structural break at observation ½ n + 1 = 35.

[99] If it may not be puzzling that a "mass indicator" is of such importance to the row sums of presumably "mass-adjusted" implied distances. Similar to Newtonian physics, the heaviest (mobile) objects are assumed to eventually gravitate towards a central position. The regions most active in FP4 will certainly be among the most central – as depicted in Figure 5, p. 93.

Estimation 1: Modelling implied distance row sums

Dependent Variable: L_IMPDISTSUM
Method: Least Squares
Included observations: 68

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.027628	0.358088	25.21064	0.0000
L_RSTAFF_TOT	-0.331632	0.033206	-9.987135	0.0000
DIST2CORE	0.000391	0.000100	3.910237	0.0002
RDXPC_GOV	-0.987681	0.296786	-3.327925	0.0015
PUBNAT_RST	-0.511642	0.120770	-4.236512	0.0001
R-squared	0.720311	Mean dependent var		5.486464
Adjusted R-squared	0.702552	S.D. dependent var		0.443151
S.E. of regression	0.241689	Akaike info criterion		0.068358
Sum squared resid	3.680062	F-statistic		40.56245
Log likelihood	2.675839	Prob(F-statistic)		0.000000

In addition to these two indicators, the two efficiency indicators *RDXPC_GOV* and *PUBNAT_RST* add explicative power. The former represents the share of government expenditure in regional R&D spending. Its performance is slightly better than that of *L_RSTAFF_GOV* (total government R&D staff) in its place, which displays a similarly negative impact. It seems that, adjusted for total research, concentration of public (non-educational) research expenditure/personnel is related to a region's centrality in the FP structure. Even more interestingly, national average natural sciences publications per researcher (*PUBNAT_RST*) induce a more central positioning. The latter variable exhibits equal values for all of a nation's regions and thus comes near to country fixed effects. The higher a country's publication output per researcher, the lower its implied distance row sums. This is mainly due to intra-country effects: First, the more "efficient" a country's research sector, the centrally it is seemingly positioned: National *PUBNAT_RST* values are weakly related to national totals of implied distances.[100] Second, we suspect peripheral regions in countries with large publication ratios to be more integrated with the national research and thus exhibit more centrality.

Since the data set included in the presented estimation only counts 68 observations, we advise caution with its parameters, despite its outstanding performance statistics. Nevertheless, its fitted values are judged as a formidable base for estimation of implied distances.

[100] National row sum totals are by construction partly dependent on the number of NUTS-1 regions per country. But even including both national totals or the national number of nodes into modelling implied distance row sums does not yield significant parameters and does not affect *PUBNAT_RST*'s performance.

6.2.2 Implied Distance Modelling

Estimated implied distance row sums (*L_IMPDISTSF*) account for core-periphery positioning in the FP cooperation structure (although they do not provide an indicator as good as obtained by adding n=68 regional dummies). Those values constitute the first indicator to be accounted for in implied distance estimation, and allow for the evaluation of bilateral distance measures on top of their effect.

In order to assess each possible distance indicator for its explicative value, we performed the stability test procedure 50 times for each double combination out of 11 distance factors.[101] The results are displayed in Table A.8 in the appendix. The factors providing the most stable model structures and at the same time exhibit high significance are *LANGSPOK*, *RSTAFFDIFF* and *RDXDIFF*.[102] Apart from these factors, *L_GEODIST* is significant in 100% of the cases, while *PATSTRUCT* is over-proportionally involved in model structures considered as stable. The factors *L_DIST2CORE*, *INDSTRUCT*, *ENGLISH* and *LANGSTUD* are clearly dominated by their counterparts.

We will concentrate on the well-performing factors and evaluate more complicated combinations of those for goodness of fit and stability. In order to determine explicative value, we will add the respective factors to an OLS model including a constant, *L_IMPDISTSF* and 68 regional dummies. The factor's exact specification (log-transformation, inverse, etc.) will be evaluated by diagnostic tests (e.g. Ramsey's RESET test, see Alexander/ Ramsey 1984).

Only if the added factor is significantly different from zero, it will be further evaluated in stability tests. This procedure already dismisses *CULTDIM*. Among the remaining factors, *RDXDIFF* and *RSTAFFDIFF* are significant, but strongly collinear, therefore only one of both may be included into the final structure. The difference in language *LANGSPOK* proves less significant than the other remaining variables. Geographical distance *GEODIST*, in contrast, exhibits large t-statistics and increases model stability when combined with other factors. However, diagnostic tests hint at misspecification of both its plain values and its log-transformed version, despite its good performance. We finally opted for its square root, a specification not rejected by Ramsey's RESET test, and providing considerable explicative

---

[101] I.e. each estimated regression had the following structure: $LL\_IMPDIST = c_0 + c_1\, L\_IMPDISTSF + c_2\, var_1 + c_3\, var_2 + \varepsilon$, where var_1 and var_2 are the respective selected distance indicators.

[102] *LANGSPOK*, however, derives its good performance seemingly from pairing with indicators of weak explanatory value.

power. The factor *PATSTRUCT* appears as a borderline case – although being significant and relatively stable, its contribution to model performance is negligible.

Apart from the cited empirical factors, we tested several dummy variables for added value. One particular proved outstanding: From the analysis in section 5.1.2 we suspected a "Romanic-Germanic divide" in node positioning, with Spanish, French, Greek and Italian regions clustering the lower right quadrant of the two-dimensional representation. In order to check for such a "Latin" cluster we constructed the dummy variable *INTRAROMANIC*, taking the value 1 if a bilateral observation involved a region speaking a Romanic language or Greek on both sides, and 0 otherwise.[103] Among the dummies tested, this dummy stands out as the only one being highly significant and supporting stability.

The final model structure contains a constant, the fitted row sum values *L_IMPDISTSF*, the square root of geographic distance *SQRT_GEODIST*, the angle between sector research expenditure vectors *RDXDIFF*, and the mentioned dummy for the distance between Romanic NUTS-1 regions *INTRAROMANIC*.[104]

Estimation 2: Implied distances modelled with bilateral factors and estimated row sums[105]

Dependent Variable: LL_IMPDIST
Method: Least Squares
Included observations: 2278

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.182888	0.064990	-33.58781	0.0000
L_IMPDISTSF	0.265045	0.006026	43.98237	0.0000
RDXDIFF	0.103716	0.020545	5.048182	0.0000
SQRT_GEODIST	0.003106	0.000278	11.16810	0.0000
INTRAROMANIC	-0.102275	0.008763	-11.67096	0.0000

R-squared	0.530828	Mean dependent var	0.847409
Adjusted R-squared	0.530003	S.D. dependent var	0.216835
S.E. of regression	0.148654	Akaike info criterion	-0.972196
Sum squared resid	50.22890	F-statistic	642.9267
Log likelihood	1112.331	Prob(F-statistic)	0.000000

Evaluating the present structure with the stability procedure outlined before results in stable estimates in 143 of the 200 simulation runs analysed.[106] Moreover, in all of the cases

[103] The following regions were classified as "Romanic": All regions being part of Spain, France, Italy, Portugal and Greece, along with BE3 (Walloon region) and BE1 (Brussels), but not Luxembourg.

[104] Adding *PATSTRUCT* to this specification results in a negligible alteration of the remaining coefficients, while the coefficient for *PATSTRUCT* is 0.061 (standard error: 0.016).

[105] Please refer to Estimation 1 for the estimation of implied distance row sums.

[106] Estimated parameter structures were classified as stable, if the Chow forecast statistic did not reject stability at a 0.01 significance level.

examined, each single factor proved significant with coefficients remaining in a relatively narrow range.

In comparison to the whole of available data, the factors in the present model are the ones that definitely cannot be dropped from the structure. Other variables may play a role but they do not deliver entirely satisfying performance.

Regarding relevance per factor, *L_IMPDIST* proves the most important.[107] The unanimously positive values yielded by this and two other factors are countered by the negative constant. The other three variables do not capture fixed effects and are therefore of lesser importance in this structure. Among those, the square root of great circle distance (*SQRT_GEODIST*) exerts the largest, positive influence, dominating compared to the artificial economic and social distance factors evaluated. Thus geography matters: the larger the physical distance between two regions, the less they are inclined to cooperate per se (i.e. adjusted for mass effects). Moreover, differences in research expenditure distribution among regional R&D sectors (as measured by *RDXDIFF*) increase implied distance and thus decrease the propensity for inter-regional collaboration. This implies that firm-dominated regions prefer to cooperate with firm-dominated regions, areas with high university participation with similar regions, etc. The cooperation among "Romanic" regions, however, seems less affected by core-periphery positioning and the cited bilateral distance factors: For each intra-Romanic pairing, the estimated value for *LL_IMPDIST* is reduced by 0.10. With current observations, this translates into an average reduction of 20% versus the value fitted from the other model factors. Thus the suspected clustering of intra-Romanic collaboration can be confirmed.

In order to determine parameter impact with regional fixed effects, we estimate the same parameters in combination with regional dummies: The coefficient for *RDXDIFF*, *SQRT_GEODIST* and *INTRAROMANIC* change slightly but significantly in size. In contrast, the factor *L_IMPDISTSF*, designed to encapsulate fixed effects, changes its sign due to the dummies' impact. This behaviour is driven by a correlation coefficient of 0.93 between *L_IMPDISTSF* and the dummies' combined impact.[108] Therefore this factor should be excluded in a model structure containing dummies. Estimating a model structure without *L_IMPDISTSF* leads to no major coefficient change for *RDXDIFF*, *SQRT_GEODIST* and *INTRAROMANIC* with respect to the model excluding dummies presented before.

---

[107] We measure impact as the arithmetic mean of the respective coefficient times the data series.

[108] The combined impact is the data series resulting from adding the respective coefficients times the dummy values, i.e. the output of the dummy part of the model structure.

Estimation 3: Implied distances modelled with bilateral distance factors and regional dummies

Dependent Variable: LL_IMPDIST
Method: Least Squares
Included observations: 2278

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.547206	0.057257	9.557083	0.0000
RDXDIFF	0.162201	0.019589	8.280040	0.0000
SQRT_GEODIST	0.004968	0.000360	13.81406	0.0000
INTRAROMANIC	-0.078640	0.011034	-7.127100	0.0000
…				

R-squared	0.691170	Mean dependent var	0.847409
Adjusted R-squared	0.681230	S.D. dependent var	0.216835
S.E. of regression	0.122424	Akaike info criterion	-1.331550
Sum squared resid	33.06295	F-statistic	69.53639
Log likelihood	1588.635	Prob(F-statistic)	0.000000

6.2.3 Explicative Power of Implied Distance Estimation

Once having determined the structure of our implied distance model, we are interested in its explanatory performance. The R^2 of 0.53 for the estimation of doubly log-transformed implied distances translates into an R^2 of 0.52 for the re-transformed "plain" values.[109]

Nevertheless, the model for implied distances explains already a considerable part of internal variation in the collaboration matrix. For illustration purposes, we simply take the collaboration matrix's row sums as a proxy for masses. In order to determine the appropriate weightings for distances and row sums, we include both factors into an OLS estimation of the (stacked and log-transformed) collaboration matrix. The constant c provides the necessary scaling factor.

Estimation 4: Modelling FP cooperation with estimated implied distances and collaboration matrix row sums

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.068533	0.228383	-35.32894	0.0000
L_IMPDISTF	-0.816150	0.037556	-21.73151	0.0000
LOG(ROWSUM+1)	0.770888	0.011061	69.69643	0.0000

R-squared	0.905152	Mean dependent var	3.763354
Adjusted R-squared	0.905071	S.D. dependent var	1.381296
S.E. of regression	0.425585	Akaike info criterion	1.130575
Sum squared resid	424.3707	F-statistic	11179.81
Log likelihood	-1323.164	Prob(F-statistic)	0.000000

[109] Those R^2 are for stacked lower triangular matrices. The corresponding values for the entire symmetric matrices are about 0.27.

The (re-transformed) fitted values exhibit a considerable R^2 of 0.93. Overall, those fitted values over-estimate the collaboration matrix. Therefore we apply the scaling procedure described in section 6.1.5. The final output thus needs no collaboration matrix information other than its row sums, but yields already an R^2 of 0.955 (or 0.92 for log-transformed values).

6.3 Modelling Implied Masses

The t-statistics of Estimation 4 underline the importance of mass indicators for modelling the FP collaboration matrix. We already mentioned different measures suitable as mass proxies, most notably matrix row sums and the first principal component. During the further course of this paper, we will rely on implied masses, the counterpart to implied distances – but keep in mind that results for this $n \times 1$ vector will be relatively interchangeable with those for other mass proxies.

However, the results of the following section should be regarded with extreme caution: The sample size of only 68 observations conforms poorly to the asymptotic properties of LS estimation. Moreover, the small sample size limits the prospects for favouring WLS over OLS, even if the "true" model may be subject to variance disparities.

6.3.1 Evaluating Eligible Factors for a Model of Implied Masses

In order to grasp an initial impression of the explicative performance of the 69 mass indicators introduced in section 5.3,[110] we evaluate R^2 and perform stability tests on each possible double combination among them according to the procedure outlined in 6.1.4. In this initial evaluation procedure, we expect ratios designed to measure efficiency to under-perform indicators introduced to describe size effects – since the latter may demonstrate their power only on top of an accurately defined model of general mass sizes.

Apart from factor doubles, the evaluation procedure uses a constant to estimate log-transformed implied masses. Stability is tested for by performing a Chow forecast test with 34 randomly drawn observations and the total of the sample – this step is repeated 30 times for each combination. The evaluation results are presented in Table A.12 in the appendix.

About one tenth of the 2346 factor doubles yielded an adjusted R^2 of more than 0.62, with stable outcomes in more than 90% of each double's 30 stability test runs. Several categories

[110] Out of the original 80 indicators, several were nearly perfectly correlated. Among those, 11 indicators were dropped that were clearly dominated by their relatives in correlation with cooperation matrix derivatives.

among the explanatory factors stand out as the best performing: First the share of regional research expenditure (or staff) in the national total, and in this respect, particularly business figures. Second, mass indicators like total research expenditure, total research staff and the number of "top" firms (by Amadeus classification). Third, human resources in research and technology, in particular with respect to the knowledge-intensive sector (KIS). One plain "efficiency" indicator figures prominently among those best-performers: *L_PATPRX_BIZ*, i.e. patents per R&D expenditure in the business sector performs particularly well in combination with the two latter categories. As expected, the values for the remaining "efficiency" indicators are considerably less impressing.

Estimation 5: Modelling implied masses

Dependent Variable: L_IMPMASS
Method: Least Squares
Included observations: 68

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.795337	0.333364	-8.385253	0.0000
L_RSTAFF_TOT	0.550476	0.037012	14.87300	0.0000
RDXPNAT_BIZ	0.699405	0.110543	6.326969	0.0000
PAT_TOPBIZ	-5.394410	0.850963	-6.339186	0.0000
DUMMY_DDR	-0.420762	0.116093	-3.624346	0.0006

R-squared	0.871911	Mean dependent var	2.362324
Adjusted R-squared	0.863779	S.D. dependent var	0.656578
S.E. of regression	0.242331	Akaike info criterion	0.073662
Sum squared resid	3.699636	F-statistic	107.2116
Log likelihood	2.495476	Prob(F-statistic)	0.000000

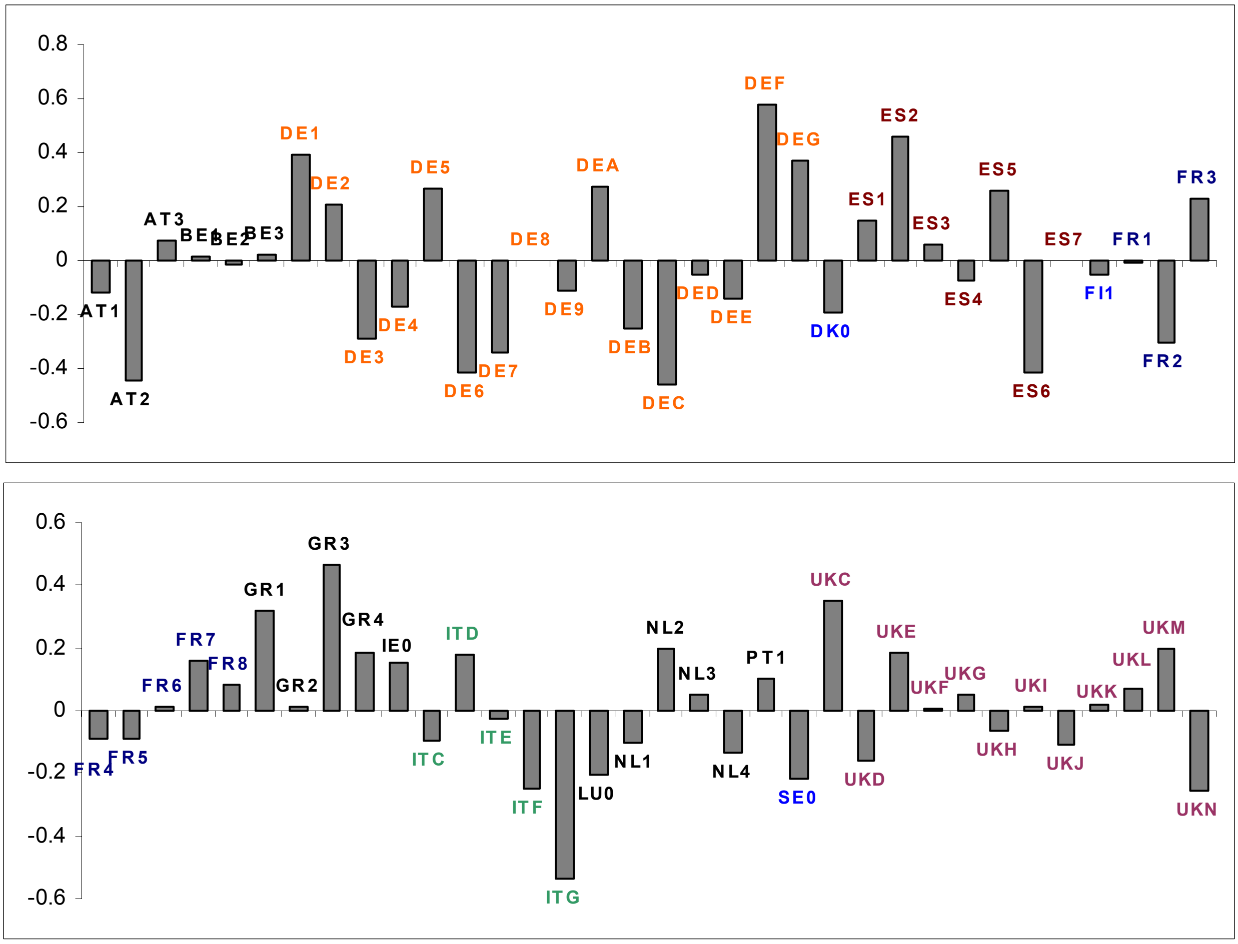
In a next step, we reduce the factor set further by identifying the dominating factors in each broad indicator category. A "dominating factor" in this case means an indicator involved in couples displaying better performance in terms of R^2, stability and t-statistics than a related factor. This greatly reduces the number of the highly correlated "plain" mass indicators, while the bilateral "dominance" among efficiency factors appears less clear-shaped. In particular, the four factors *L_RSTAFF_TOT*, *L_RDX_MPTOT*, *L_KIS_HRSTC* and *L_TOPBIZ* stand out among plain mass indicators, and *L_RDXPNAT_BIZ* and *L_RSTAPNAT_BIZ* in the category "regional share in national total". A further evaluation of those four and related "plain" mass indicators in combination with efficiency ratios reveals that the combination of *L_RSTAFF_TOT* and *L_RDXPNAT_BIZ* performs best.[111] With these two basic factors, especially patent-related efficiency ratios perform well (i.e. patents per researcher, per researcher in the business sector, per employed person, per firm etc), put in particular *L_PAT_TOPBIZ* (patents per top firm) stands out – and outperform *L_PATPRX_BIZ*.

[111] The factor *L_RSTAFF_TOT* is equivalent to *L_RDX_MPTOT*. Similarly, *L_RDXPNAT_BIZ* and *L_RSTAPNAT_BIZ* are virtually interchangeable. Of their four possible pairings, the combination *L_RSTAFF_TOT* and *L_RDXPNAT_BIZ* performs marginally better.

Interestingly, all of these indicators, and even the total number of patents, exhibit a significant negative coefficient, hinting at an inverse relationship between implied mass and patent applications.

Further potential for model enhancement is found by analysing the structure's residuals: In particular, the five German ex-GDR *Länder* are thoroughly over-estimated by nearly every model structure tested up to this point. Therefore we introduce a "GDR" dummy variable (*DUMMY_DDR*), which adds explicative value and remains robust in stability test procedures.

Figure 10: Residuals of implied mass estimation (Estimation 5)[112]



The final model structure for estimating log-transformed implied masses (*L_IMPMASS*) thus contains the log-transformed number of research staff (*L_RSTAFF_TOT*, positive impact), a variable representing the regional share in national business research expenditure

112 Note: negative residuals correspond to over-estimated values, positive ones to under-estimation.

(*L_RDXPNAT*) and a factor setting total patent applications in relation to the number of top firms (*L_PAT_TOPBIZ*).[113] While the two former factors conform to expectations, the latter puzzles in its impact and significance, and by the fact that it outperforms the separate inclusion of *L_TOPBIZ* and *L_PAT_M*. In addition, the area of the former GDR is subject to substantial negative fixed effects. Further interpretation of those results will be given in section 7.

The performance statistics of Estimation 5 indicate satisfying explicative power and diagnostic tests do not hint at misspecification.[114] Having found the final model structure, we once again include the remaining data set factors one by one and test whether they increase performance. As expected, none of those variables enhances the presented model structure.[115]

Figure 10 displays the residuals of implied mass estimation as obtained by Estimation 5: positive values correspond to over-estimated values, while negative ones represent under-estimation. It is difficult to identify general regional patterns based on these ex-GDR adjusted residuals, though several values appear notable: First, the parameters of Estimation 5 seem to thoroughly overstate masses in the Italian and Spanish south. Fitted values for large-mass regions like BE1, FR1, ITC or UKI do not deviate much from the observed figures, while the importance of German, Spanish, Dutch and particularly Greek economic core regions appears undervalued. Interestingly, residuals for Scandinavian countries seem rather be biased to the negative, while the single-region countries Ireland and Portugal display positive values. Due to the small sample size of Estimation 5 these results should be treated with prudence, and further interpretation will be carried out in the light of other results in section 7.

---

[113] Although a constant was included in the model structures initially evaluated, it proved fairly insignificant and was therefore dropped from the final structure.

[114] Apart from the indicators displayed, Estimation 5 displays normally distributed residuals by the Jarque-Bera statistic, is not rejected by the RESET test and its Theil covariance proportion (Greene 2003) surpasses 0.90. We attempted to decrease the Theil variance proportion by WLS estimation with model factors, but were not able to reduce this slight disadvantage in estimator variance. Thus we deem OLS Estimation 5 not to be mis-specified under the given circumstances.

[115] This final procedure is oriented at the one described in section 6.1.4. Each of the remaining factors was added to the structure in Estimation 5 and the resulting structure's consistency was evaluated in 100 stability test runs. Moreover, the number of significant t-statistics per randomly drawn sub-sample provided information on a factor's added value. The most significant, *L_KIS_HRSTC*, exhibited significant t-statistics in 25% of the cases, compared to 80% for the weakest member of Estimation 5, *DUMMY_DDR*.

6.3.2 Explicative Power of the Models for Implied Masses And Implied Distances

Once having determined a model structure for implied distances (Estimation 2, without regional dummies) and implied masses (Estimation 5), we are able to combine both in order to obtain an estimate for the FP collaboration matrix. Equation (27) describes the necessary model structure as multiplicative, without the need for a scaling factor θ.

$$T_{ij} = \frac{M_i^* M_j^*}{D_{ij}^*}$$

Most elements in the 2346x1 vector of estimated implied distances $\hat{\mathbf{D}}^*$ stem from Estimation 2, while the observations corresponding to the implied distance matrix main diagonal elements were set to 1 by construction. Estimation 5 yields a 68 x 1 vector of estimated implied masses $\hat{\mathbf{M}}^*$. By constructing the outer product of the latter we obtain a 68 x 68 matrix containing elements $\hat{M}_i^* \hat{M}_j^*$, and its stacked 2346x1 vector $\hat{\mathbf{M}}^*$.

We divide the elements of $\hat{\mathbf{M}}^*$ by the corresponding elements of $\hat{\mathbf{D}}^*$, which yields an estimate of FP collaboration exhibiting an R^2 to actual FP collaboration of 0.85 (in stacked form). Scaling this output by the procedure described in section 6.1.5 considerably enhances the fit and drives R^2 to 0.953.[116] This fit is equivalent to the estimation based solely on row sums, but in contrast provides additional insight into the nature of regional "masses" in the FP context.

In order to adjust for scale misfits, we tried to combine the estimates for log-transformed implied masses and implied distances in an OLS estimation (including a constant) rather then by a simple division. However, this approach yielded only a negligible increase in the R^2 of log-transformed output, and performed even worse with plain values compared to the algebraic combination outlined before.

We have broken down the absolute and relative distribution of FP collaboration links into three distinct dimensions:

- First, node-specific mass effects appear to be driven by the total number of research staff, and the importance of regional business R&D versus the national total. The number of top firms evenly displays positive impact on cooperation, while patents

[116] In log-transformed equivalents, R^2 of the combined structure is 0.85 and of the scaled estimation output 0.92.

interestingly affect “mass” negatively. During 1994-1998, the eastern German *Länder* apparently exhibited less “mass” size compared to those factors than the rest of the EU-15.

- Second, another node-specific factor is found in core-periphery positioning in the FP continuum: The more distant a region is to the national economic core, the less centrally it is positioned in FP collaboration. Again, total research staff plays a role in rendering regions more central and thus less biased in the distribution of collaborative links. Similarly, regions in member states with high publication output and disposing of important governmental research position themselves in the centre of regional FP links.
- Third, bilateral distances, i.e. relative impediments to bilateral collaboration, depend positively on geographic distance. Moreover, differences in the sector distribution of research expenditure impede cooperation between regions. Both factors matter less to collaboration between Romanic and Greek language regions, which display a certain clustering among them.

6.4 Modelling FP Collaboration Directly

In order to assess the validity of the results obtained by the estimation of implied masses and implied distances, the evaluation procedure laid out in section 6.1.4 will be applied to direct estimation of FP collaboration data. This approach allows for an intermingling of mass and distance factors, and benefits from the advantages associated with large sample size.

6.4.1 Variance Adjustment Considerations

As has been mentioned in section 6.1.2, a model of the collaboration matrix may be subject to residuals not independently and identically distributed. We concluded that node-specific deviations from a residual expected value of 0 could easily be corrected for by regional dummy variables.

The variance structure poses more problems: All of the model structures to be mentioned in this section were rejected by White's heteroskedasticity test. This may either be due to heteroskedasticity or to incorrect specification. Since our study is exploratory and since we do not dispose of an unlimited data set, the latter case cannot be ruled out.

In order to obtain a specific variance model, WLS estimation with the options described in section 6.1.2 were evaluated, but none of those possibilities yielded satisfying results. Moreover we checked the correlations of the squared residuals of the best-performing model structures against each of the eligible factors in our data set. The most promising of those prospective weighting series exhibited an R^2 with squared residuals of roughly 0.06 and conformingly did not add any value. Therefore the presentation of WLS results will be omitted hereafter.

6.4.2 Reducing the Set of Eligible Explanatory Variables

The purpose of this sub-section is to evaluate whether an alternative modelling approach yields results similar to those obtained by the investigation of implied masses and distances. Therefore the according procedure is intended to restrict influence by results of sections 6.2 and 6.3 as far as possible.

The model finding procedure starts with the specification of the dependent variables. Absolute cooperation numbers are roughly exponentially distributed, therefore we take their log-transformation to achieve a near-to-normal distribution. The resulting factor *L_COOPFP4*, however is not sufficiently smoothly distributed over the total sample to satisfy standard normality test requirements – it is slightly skewed to the right (-0.25) and its kurtosis (2.78) is below the normality requirement. But neither a second log-transformation nor powering by exponents of <1 yield any noteworthy improvement in distribution. Consequently we henceforth employ *L_COOPFP4* as the dependent variable.

Since FP collaboration is once again modelled "from scratch", we evaluate the quasi-entirety of the eligible factor set by the approach outlined in section 6.1.4. Due to the corresponding routine's massive demand for computing power, the first task consists of excluding possibly redundant, dominated variables. Similar to section, we use correlation among factors and with the dependent variable to single out underperforming factors: Among the collinear "mass" factors measured in money terms, correlations favour PPP data over corresponding series in ECU/Euro. Moreover, several macroeconomic and human capital (HRST) factors are not included in the sample. The quantities for "top" firms with more than 200 employees are excluded since they are mostly identical with the number of "normal" firms. This leaves us with a set of 58, mainly log-transformed "mass" factors.

In addition to mass effects, a "straight" collaboration model ought to incorporate distance factors. Hence all of the 8 variables introduced in section 5.2 are evaluated in absolute terms or log-transformation according to the one of both specifications that is nearer to normal distribution. From our experience with implied distance estimation we take the square roots rather than the logs of *GEODIST* and further introduce *L_DIST2CORE* (see p. 126). Thus we evaluate all possible pairs of 67 variables (along with a constant) for model consistency in estimating *L_COOPFP4*. In order to keep computing time in check, we chose to run the evaluation routine for the relatively low number of 50 random sub-sample pairings for each variable double. For illustration purposes Table A.13 in the appendix provides a fragment of evaluation results.

Each variable pair is checked for significance and consistency of t-statistics, for the R^2 achieved over the whole sample, and for the number of insignificant (i.e. consistent) Chow F-statistics. Among the results for each factor, total research staff *L_RSTAFF_TOT* stands out with the highest R^2 and significant (positive) t-statistics in 100% of the cases. Moreover, its 66 pairings perform reasonably well in Chow forecast tests.[117]

Along with *L_RSTAFF_TOT*, but with slightly inferior results, rank the four KIS HRST factors[118], total research expenditure *L_RDX_MPTOT* and total GDP (*L_GDP_MP*). In

---

[117] In sum, variable pairs with *L_RSTAFF_TOT* yielded consistent model structures (as measured by the Chow forecast statistic) in 66.6% of all cases. However, a considerable number of pairs exhibit consistent structures in more than 90% of the cases, mainly due to their factors' indecisive impact on the dependent variable.

[118] Knowledge intensive services human capital in science & technology: *L_KIS_HRST*, *L_KIS_HRSTC*, *L_KIS_HRSTE*, *L_KIS_HRSTO*.

addition, the number of "top" firms *L_TOPBIZ*, respectively *L_NBBIZ10* (the number of firms with more than 1000 employees), yields comparable results, particularly in model consistency. The corresponding t-statistics for those factors are thoroughly positive in nearly all of the cases. Among "efficiency" indicators, the group of "regional share in national total" indicators for research expenditure per sector deliver once again exceptionally good results: All of them boast high R^2 and in particular *L_RDXPNAT_EDU*[119] exhibits superior Chow test performance. Moreover, two factors describing the share of research staff in the regional labour force (*RSTAFF_PCLTOT* and *RSTAFF_PCLGOV*) display noteworthy F-statistics and R^2 along with stable t-statistics.

In addition, the following efficiency ratios exhibit unanimous t-statistics along with medium R^2 and Chow test results: Publications per researcher (*PUBNAT_RST*, positive), patents per research expenditure (*L_PATPRX_BIZ*, negative), expenditure per researcher in the business sector (*XPRES_BIZ*, positive) – or in the education or government sector[120] - and GDP per employee in PPP (*L_GDPP_EMP*, positive). Interestingly, patents per top firm (*L_PAT_TOPBIZ*), the indicator of importance in implied mass estimation, is involved in structures with medium R^2, but displays absolutely unconvincing t-statistics.

Among distance indicators, several combinations of *SQRT_GEODIST*, *PATSTRUCT* and *LANGSPOK* stand out in consistency performance, while *L_RDXDIFF* (resp. *L_RSTAFFDIFF*) along with *L_DIST2CORE* exhibit higher R^2 with a lower share of consistent structures. The corresponding t-statistics are entirely negative and significant in 85% to 100% of the cases.

An analysis of test results for specific pairings among the cited factors allows for dropping several among them due to inferior performance with respect to collinearity: However, a considerable number of eligible factors remains. In particular, the total number of research staff (*L_RSTAFF_TOT*) performs better in comparison to sector-specific staff, to R&D expenditure and to total GDP. Moreover, *LANGSPOK* seems slightly inferior to the other distance factors

6.4.3 Finding A "Final" Model Structure

The final steps for the selection of "final" model factors consist of intensely evaluating the most promising combinations of the reduced factor set for stability and goodness of fit. In a

---

119 Regional research expenditure in educational institutions as a share of national total educational research expenditure

120 Interestingly, the „expenditure per research" factors *XPRES_BIZ*, *XPRES_EDU* and *XPRES_GOV* differ in their behaviour.

first approach, the dependent variable *L_COOPFP4* is regressed on the entire set of remaining "non-collinear" factors. The added value of variables deemed as "collinear" is tested for alternately, i.e. either one or the other factor is added to an otherwise equal model structure.

These general factors are complemented by 68 regional dummy variables in order to evaluate their explanatory power on top of node-specific fixed effects. Since those dummies inhibit us from applying the stability procedure as has been outlined in section 6.1.3, we use the fitted values from the implied distance row sum estimation Estimation 2 instead of regional dummies. This allows for the section 6.1.3 procedure, which enables us to evaluate factor stability in random subsets.

Estimation 6 displays one of those prior estimations over the sample of bilateral connections among 34 randomly drawn regions.

Estimation 6: testing structures for modelling FP collaboration directly

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 595

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_1_ONESER	40.62345	12.72337	3.192823	0.0015
_1_L_RSTAFF_TOT	0.369467	0.049430	7.474580	0.0000
_1_L_RDXDIFF	-1.252994	0.205890	-6.085760	0.0000
_1_LL_RDXPNAT_G	0.112396	0.019696	5.706483	0.0000
_1_L_IMPDISTSF	-1.068144	0.122670	-8.707433	0.0000
_1_L_DIST2CORE	-0.008810	0.006290	-1.400745	0.1618
_1_SQRT_GEODIST	-0.014014	0.002310	-6.065808	0.0000
_1_XPRES_GOV	-30.71690	11.82833	-2.596893	0.0096
_1_PATSTRUCT	-0.530223	0.133005	-3.986492	0.0001
_1_L_PAT_TOPBIZ	-0.101801	0.019957	-5.101097	0.0000
_1_PUBNAT_RST	0.235507	0.152577	1.543531	0.1232

R-squared	0.828909	Mean dependent var	3.600954
Adjusted R-squared	0.825979	S.D. dependent var	1.356598
S.E. of regression	0.565915	Akaike info criterion	1.717570
Sum squared resid	187.0320	F-statistic	282.9389
Log likelihood	-499.9771	Prob(F-statistic)	0.000000

The various test results allow for the following general statement: factors with a t-statistic of less than 5 in a total-sample estimation are unlikely to survive evaluation according to the stability check procedure. The model structure in Estimation 6, for instance, exhibits insignificant Chow forecast test-statistics (i.e. consistency) in 131 out of 200 test runs. In particular, the apparently significant factors *XPRES_GOV* and *PATSTRUCT* display insignificant t-statistics in several of the random subsets, and show a relatively wide distribution of OLS coefficients over the 200 runs. Excluding those variables from the structure above improves the share of insignificant Chow stats to 142 out of 200.

In order to adjust for specific fixed effects, several dummy variables are added to the general mass and distance factors. Most of them are subsequently quickly excluded for unsatisfying performance, but several dummies seem to have an impact: The country dummies for

Greece and Spain and the dummy for objective-1 regions[121] provide some added value but perform poorly in the consistency test. The "Intra-Romanic" dummy and the dummy for the Eastern German *Länder* (*DUMMY_DDR*), in contrast, add considerable explanatory value with high t-stats and exhibit convincing consistency test results. We omit the former dummies, but keep the latter since they enhance the performance of the remaining general factors.

Out of the general factors cited above, the factors capturing research staff in comparison to the labour force (*L_RSTAFF_PCLTOT* and *L_RSTAFF_PCLGOV*) exhibit relatively low t-stats in an estimation over the entire sample and perform even more poorly with the consistency test procedure.

The decision is harder for the variables describing sector-specific expenditure per researcher *XPRES_BIZ*, *XPRES_EDU* and *XPRES_GOV*. Tested together with the remaining eligible factors, *XPRES_GOV* performs fairly well, but tears down overall consistency check results. Therefore it is left out in the final structure, although by classic tests it should definitely be included in the structure.

In contrast, the entirety of factors describing the regional share in total national expenditure performs well, but in combination with the remaining factors the one for total regional R&D expenditure (*LL_RDXPNAT_T*) proves best with regressions statistics and in the consistency test.

The "distance" factor *PATSTRUCT* exhibits a t-statistic of about 5, but does not add sufficient value to overall consistency test results; in addition it does not provide entirely satisfying individual test results. Natural science publications per researcher (*L_PUBNAT_RST*) corresponds to relatively high t-stats but does less well in the consistency check, in particular for the sample confined to Spanish, French and British regions. The results for both *L_PUBNAT_RST* and *L_PAT_TOPBIZ* are considerably lowered when the dummies *DUMMY_DDR* and *INTRAROMANIC* are added to the model structure. However *L_PAT_TOPBIZ* still adds value to the overall structure and its impact on consistency test performance is none, i.e. decisive neither in the one nor in the other direction.

The factors for bilateral geographic distance (a "distance" factor) and for distance to the economic core (a "mass" factor by construction) add explicative value, but underperform in

---

[121] Those NUTS-1 regions are defined as „objective-1 regions", where more than half of the territory's population belongs to a region declared as objective-1 by the European Commission during 1994-1998 (Ecotec 2003, p. 55). The corresponding dummy displays 1 if an „objective-1 region" is part of a collaboration matrix observation, and 0 otherwise.

the consistency test when added solely to the remaining variables. A combination of both, in contrast, performs better in testing for consistency. In this case, the log-transformed specification of geographic distance *L_GEODIST* performs slightly better than its square root *SQRT_GEODIST* previously employed in implied mass estimation.

On basis of the considerations above, we confine the “final” model structure to the following factors:

- *C*: a constant
- *L_RSTAFF_TOT*: (log-transformed) total research staff in full-time equivalents
- *LL_RDXPNAT_T*: the regional share in total national R&D expenditure, aiming at measuring a region’s importance versus the national level
- *L_RDXDIFF*: an indicator measuring bilateral disparity in (log-transformed) R&D expenditure distribution over sectors.
- *PATSTRUCT:* An indicator measuring bilateral disparity in sector distribution of EPO patent filings.
- *INTRAROMANIC*: A dummy of value one if an observation corresponds to collaboration between two Romanic-/Greek-language regions
- *DUMMY_DDR*: A dummy of value one for those observations where one of the five new German *Länder* is involved
- *L_IMPDISTSF*: The fitted values obtained from the estimation of (log-transformed) implied distance row sums, employed as a general proxy for region-specific fixed effects

Estimation 7: “Final“ structure for direct FP collaboration estimation

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	17.01249	0.729066	23.33464	0.0000
L_RSTAFF_TOT	0.171489	0.016040	10.69144	0.0000
LL_RDXPNAT_T	0.148137	0.008211	18.04186	0.0000
L_RDXDIFF	-1.140216	0.091768	-12.42493	0.0000
PATSTRUCT	-0.488636	0.056681	-8.620741	0.0000
INTRAROMANIC	0.586992	0.031261	18.77729	0.0000
DUMMY_DDR	-0.485566	0.035206	-13.79197	0.0000
L_IMPDISTSF	-1.354910	0.043334	-31.26633	0.0000

R-squared	0.853134	Mean dependent var	3.763354
Adjusted R-squared	0.852694	S.D. dependent var	1.381296
S.E. of regression	0.530147	Akaike info criterion	1.572079
Sum squared resid	657.1082	F-statistic	1940.185
Log likelihood	-1836.048	Prob(F-statistic)	0.000000

The structure from Estimation 7 exhibits insignificant Chow forecast statistics (i.e. "consistency") in about two thirds of randomly determined sample divisions. In 100% of 1000 trials, the factors' t-statistics proved significant and pointed unanimously into one direction. [122]

Apart from Estimation 7 members, several variables could be included since each of them raises the estimation's R^2 by about one point and since their inclusion is supported by standard tests such as the Wald test (Greene 2003, p. 164). However, they generally tear down overall consistency results more than the factors already included do. Therefore they are left out in the presented final structure. When examined in the consistency check procedure, all of these factors exhibit unanimous t-statistics, but over the sub-samples analysed their coefficients vary more than those of the "final" factors presented above. Thus these factors matter apparently, but their impact depends on the sub-sample to be estimated, rather than being equal over the whole sample.

This variable set is constituted as follows – while the factors not mentioned hereafter did prove important enough to improve the "final" structure's fit.[123]

- *SQRT_GEODIST*: (square root of) great circle distance between regions' geographic centres. Negative impact, particularly in combination with l_dist2core.
- *L_DIST2CORE*: (log-transformed) great circle distance to the geographic centre of a country's economic core (the region with largest GDP on the national level). Negative impact.
- *LL_XPRES_GOV*: (doubly log-transformed) governmental R&D expenditure per researcher in this sector. The only clearly significant factor among the "expenditure per researcher" factors. While *LL_XPRES_EDU* and *LL_XPRES_BIZ* prove insignificant, *LL_XPRES_GOV* exerts negative influence over nearly all sub-samples, but the size of its impact varies considerably from subset to subset.

The following factors would enhance R^2 etc. but exhibit high collinearity with *L_RSTAFF_TOT* and thus would decrease consistency test results.

- *L_TOPBIZ*: (log-transformed) number of businesses ranking in the "Amadeus Top 1,500,00". Positive impact, but insignificant for several sub-samples.

---

[122] In 1000 runs, the Chow forecast statistic over two randomly drawn sub-samples proved insignificant in 650 cases. The 100% significant and unanimous t-statistics refer to the estimation over the two randomly drawn sub-samples, i.e. over m (m-1) /2 + m (m-1) /2 = 595 + 595 = 1190.

[123] I.e. they displayed either insignificant or fairly low significant t-statistics.

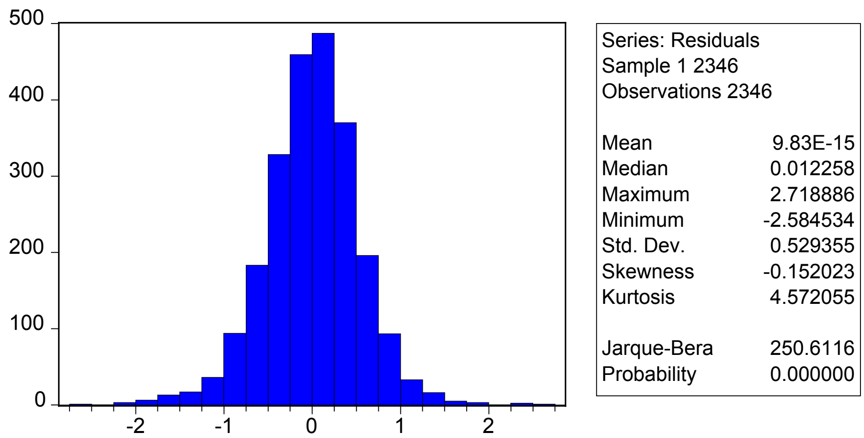
- *L_TOT_HRSTC*: (log-transformed) total human resources in Science & Technology (as defined by Eurostat). Positive impact.
- *L_RDX_MPTOT*: (log-transformed) amount of total regional R&D expenditure. Positive impact.

6.4.4 Properties of the "Final" Model Structure

In order to ensure the proper interpretation of standard methods such as t-statistics, an OLS model's residuals should at best follow normal distribution. (Besides, a linear model is in practice more likely to achieve a normal distribution of residuals, if the explanatory, but in particular the dependent variable are about normally distributed.)

Figure 11 displays the histogram for the "final" structure's residuals, which confirms that residuals are not normally distributed. However the present residual distribution constitutes one of the nearest-to-normal outcomes feasible with the present data structure.

Figure 11: Histogram of Estimation 7's residuals



Standard diagnostic tests on the structure, in particular Ramsey's RESET test and Chow's heteroskedasticity test, return mixed results: The structure is rejected by higher-order versions of the former test and the latter test points to the presence of heteroskedasticity. This is mainly due to regional fixed effects, and the heteroskedasticity is alleviated to some extent by adding 68 regional dummies to the structure (Estimation 8).

Prior to the analysis of regional fixed effects, we assess the interconnectedness of Estimation 7's regressors. The variables' correlation matrix over the entire sample is presented in Table 9.

Table 9: Correlation matrix of the regressors in Estimation 7

	Mass factor	Mass factor	Distance factor	Distance factor	Dummy	Dummy	Mass factor
	L_RSTAFF_TOT	LL_RDXPNAT_T	L_RDXDIFF	PATSTRUCT	INTRAROMANIC	DUMMY_DDR	L_IMPDISTSF
L_RSTAFF_TOT	1.00	0.50	-0.27	-0.29	-0.11	-0.16	-0.86
LL_RDXPNAT_T	0.50	1.00	-0.10	0.04	0.11	-0.42	-0.59
L_RDXDIFF	-0.27	-0.10	1.00	0.35	0.11	0.07	0.25
PATSTRUCT	-0.29	0.04	0.35	1.00	0.02	-0.02	0.28
INTRAROMANIC	-0.11	0.11	0.11	0.02	1.00	-0.18	0.11
DUMMY_DDR	-0.16	-0.42	0.07	-0.02	-0.18	1.00	0.21
L_IMPDISTSF	-0.86	-0.59	0.25	0.28	0.11	0.21	1.00

Although many coefficients are far from zero, most of them seem good enough from an econometrics perspective. However, the correlation between "mass" factors is disturbingly elevated. The correlation coefficients between *L_IMPDISTSF* and the other mass factors (-0.86 and –0.59 respectively) stem from the fact that *L_IMPDISTSF* was introduced for representing regional fixed effects, which in turn constitute an important element of "masses". But since regional dummies will replace this factor later on (Estimation 9), those considerations are of lesser relevance. The correlation between *L_RSTAFF_TOT* and *LL_RDXPNAT_T* (0.50) is not negligible, but it has to be noted that this correlation is already minor compared to those among other mass factors.

In addition to regressor collinearity, we are as well interested in parameter structure stability. For that purpose, the procedure described in section 6.1.3 has been applied 1,000 times: For each of the 1,000 runs, a set of m=34 nodes out of the n=68 in total is randomly selected and a second set of equal size is constructed out of the remaining nodes. Subsequently, two intra-set collaboration matrices with m (m-1) /2 elements each are compiled and the model structure of Estimation 7 is estimated on both subsets. In order to determine whether there a structural break between both subsets exists, a Chow forecast F-statistic is computed by comparing a model estimated over both subsets ("combining regression") with the fit of the model for the first subset. If the Chow statistics confirm the null hypothesis of no structural break, the model structure is deemed as consistent. Moreover the combining regression's R2, coefficients and t-statistics are stored.

In the case of the structure in Estimation 7, about 650 runs out of 1000 imply consistency. Although the remaining 350 trials were rejected by Chow's forecast statistic, the corresponding estimation results turned out rather similar to the "consistent" ones: In the 1000 combining regressions evaluated, parameter estimates were either positively or negatively significant in 100% of the cases. Moreover the relative dimensions of coefficients remain about equal, when comparing the 5%-confidence interval upper and lower bounds[124]

[124] If coefficient estimates are assumed to be normally distributed, one may define a bandwidth for a parameter estimate with a Type I Error probability of 5% as follows: coefficient estimate plus (minus)

for coefficients: In all cases, the constant proved the largest parameter and in all cases, its lower bound was larger than the upper bound of the next smaller parameter, namely the *INTRAROMANIC* dummy factor. The factor *INTRAROMANIC*, in turn, was in 100% of the cases larger than both of the "plain" mass indicators *L_RSTAFF_TOT* and *LL_RDXPNAT_T*.

The two negative factors the nearest to zero were *PATSTRUCT* and *DUMMY_DDR*. Both factors display coefficients of about equal size. In all cases, those coefficients were higher than the ones of *L_RDXDIFF* and *L_IMPDISTSF*, which both proved fairly negative.

Table 10: Coefficient values from consistency test runs[125]

	C	L_RSTAFF_TOT	LL_RDXPNAT_T	L_RDXDIFF	PATSTRUCT	INTRAROMANIC	DUMMY_DDR	L_IMPDISTSF
Max. upper bound	20.40	0.25	0.19	-0.77	-0.34	0.74	-0.30	-1.09
Median coef	16.61	0.17	0.15	-1.24	-0.58	0.59	-0.48	-1.30
Min. lower bound	12.95	0.09	0.11	-1.75	-0.83	0.45	-0.69	-1.54
Rank	1	3	3	7	5	2	5	7

When applying the same methods not to the results of the combining regression (1190 observations), but to subset estimation (595 observations), the results given above are confirmed, although with less outstanding results: Over all runs, t-statistics for most parameters were significant in 100% of the cases – only *L_RSTAFF_TOT* proved a bit inferior with significant t-statistics only in 88% of all cases. The relative size of coefficients presented in Table 10 remains the same, although a decisive boundary cannot be drawn between negative coefficients *PATSTRUCT* and *DUMMY_DDR* on the one side and *L_RDXDIFF* and *L_IMPDISTSF* on the other.

Concluding, we deem the present structure as sufficiently stable and as representing a sufficient approximation of dependent data. We have determined the sign and relevance of factors, as well as the relative structure of model parameters. However, the present structure may not deliver a precise measure of the factors' impact.

In order to assess regional specificities, we try to account for fixed effects by including regional dummies. If the structure of Estimation 7 is sufficiently general, the parameter estimates with regional dummies should remain equally significant. Of course, this does not apply to *L_IMPDISTSF* (our proxy for fixed effects). Moreover coefficient estimates should

1.96 times its standard error yields the upper (lower) bound. The method described, compares the number of times a parameter's lower bound is greater than the next smaller parameter's upper bound.

[125] The coefficients displayed are the median of 1000 random sub-sample estimations. The minimum lower bound is the minimum of the 1000 coefficients minus 1.96 times the respective standard error. The maximum upper bound is obtained likewise.

remain about the same as in Estimation 7, except for the constant, which will certainly be altered by the inclusion of 68 “node-specific constants”.

Estimation 8: “Final“ structure for direct FP estimation with regional fixed effects

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.643878	4.576485	0.140693	0.8881
L_RSTAFF_TOT	0.220894	0.100238	2.203706	0.0276
LL_RDXPNAT_T	0.150697	0.046575	3.235586	0.0012
L_RDXDIFF	-0.747315	0.082914	-9.013141	0.0000
PATSTRUCT	-0.339480	0.061169	-5.549889	0.0000
INTRAROMANIC	0.393110	0.034273	11.46982	0.0000
DUMMY_DDR	-0.622515	0.133841	-4.651151	0.0000
L_IMPDISTSF	0.120107	0.267677	0.448700	0.6537
…				
…				

R-squared	0.920082	Mean dependent var	3.763354
Adjusted R-squared	0.917441	S.D. dependent var	1.381296
S.E. of regression	0.396889	Akaike info criterion	1.021537
Sum squared resid	357.5716	F-statistic	348.4529
Log likelihood	-1122.263	Prob(F-statistic)	0.000000

Estimation 8 presents the effect of 68 regional dummies (not displayed) on the “final” regressors. As expected, the inclusion of regional dummies renders *L_IMPDISTSF* redundant. The constant is reduced, even if the sum of regional dummy coefficients is negative. Consequently, most of the general coefficients in Estimation 8 turn out higher than their counterparts in Estimation 7. Moreover, the constant is insignificant which is apparently due to the now futile inclusion of *L_IMPDISTSF*. Therefore we exclude the latter factor from the Estimation 8’s structure and obtain Estimation 9.

The exclusion of *L_IMPDISTSF* renders the constant significantly positive, albeit at a low margin. Most of the other coefficients do not change noticeably versus Estimation 8. Interestingly, the “mass” factors’ coefficients are nearly the same as in Estimation 7, although regional dummies were designed to capture “mass” effects. However, they considerably reduce the impact of “distance” factors. Therefore we suspect regional dummies to rather describe core-periphery effects than mass effects. The supposition is confirmed by a look at dummy coefficient’s correlations to Estimation 5’s residuals (from modelling implied masses) and regional dummy coefficients in Estimation 3 (modelling implied distances): While the former exhibit a correlation coefficient of 0.44 with Estimation 9 dummy parameters, the latter display a correlation of 0.90.

Evenly noteworthy, the dummy variables *INTRAROMANIC* and *DUMMY_DDR* withstand regional dummies and prove still significant. Though, the t-statistics of all general indicators are considerably reduced in comparison to Estimation 7. Thus even if the factors employed in the “final” structure may be regarded as general, their relevance seems of far lesser importance than that of regional dummies. This may either imply that substantial node-specific effects are still not accurately accounted for, or that nodes' behaviour differs too much to be encapsulated in a model consisting only of general factors.

Estimation 9: "Final" structure for direct FP estimation with regional fixed effects without log-transformed implied distance row sum fit *L_IMPDISTSF*

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.620013	1.243898	2.106292	0.0353	…				
L_RSTAFF_TOT	0.184519	0.058944	3.130423	0.0018	DUMMY_FR1	0.537918	0.129556	4.152025	0.0000
LL_RDXPNAT_T	0.143093	0.043375	3.298992	0.0010	DUMMY_FR2	-0.725919	0.082317	-8.818542	0.0000
L_RDXDIFF	-0.747428	0.082899	-9.016130	0.0000	DUMMY_FR3	-0.723841	0.112101	-6.457051	0.0000
PATSTRUCT	-0.340281	0.061132	-5.566316	0.0000	DUMMY_FR4	-0.819073	0.080632	-10.15813	0.0000
INTRAROMANIC	0.393572	0.034252	11.49048	0.0000	DUMMY_FR5	-0.634401	0.081337	-7.799626	0.0000
DUMMY_DDR	-0.622216	0.133816	-4.649791	0.0000	DUMMY_FR6	-0.462607	0.077396	-5.977154	0.0000
DUMMY_AT1	-0.572436	0.103473	-5.532202	0.0000	DUMMY_FR7	-0.245289	0.084200	-2.913180	0.0036
DUMMY_AT2	-1.424738	0.104880	-13.58451	0.0000	DUMMY_FR8	-0.361804	0.079838	-4.531720	0.0000
DUMMY_AT3	-1.086230	0.103018	-10.54412	0.0000	DUMMY_GR1	-0.827130	0.127916	-6.466189	0.0000
DUMMY_BE1	-0.434152	0.098610	-4.402729	0.0000	DUMMY_GR2	-1.118541	0.147360	-7.590520	0.0000
DUMMY_BE2	0.023830	0.102962	0.231444	0.8170	DUMMY_GR3	0.167096	0.120310	1.388876	0.1650
DUMMY_BE3	-0.511320	0.099939	-5.116314	0.0000	DUMMY_GR4	-0.870039	0.154525	-5.630427	0.0000
DUMMY_DE1	0.140591	0.124285	1.131198	0.2581	DUMMY_IE0	-0.015386	0.138410	-0.111159	0.9115
DUMMY_DE2	0.086697	0.123444	0.702316	0.4826	DUMMY_ITC	0.309074	0.103144	2.996517	0.0028
DUMMY_DE3	-0.537076	0.096433	-5.569406	0.0000	DUMMY_ITD	-0.135356	0.080500	-1.681450	0.0928
DUMMY_DE4	-0.281178	0.165384	-1.700158	0.0892	DUMMY_ITE	0.017879	0.095076	0.188050	0.8509
DUMMY_DE5	-0.596816	0.108375	-5.506973	0.0000	DUMMY_ITF	-0.676753	0.082281	-8.224885	0.0000
DUMMY_DE6	-0.714521	0.090455	-7.899227	0.0000	DUMMY_ITG	-1.444637	0.094702	-15.25456	0.0000
DUMMY_DE7	-0.712011	0.097588	-7.296121	0.0000	DUMMY_LU0	-2.173265	0.188484	-11.53022	0.0000
DUMMY_DE8	-1.174766	0.179602	-6.540941	0.0000	DUMMY_NL1	-1.265385	0.093969	-13.46601	0.0000
DUMMY_DE9	-0.290835	0.095201	-3.054949	0.0023	DUMMY_NL2	-0.053591	0.079477	-0.674299	0.5002
DUMMY_DEA	0.203175	0.116987	1.736734	0.0826	DUMMY_NL3	0.440902	0.101950	4.324680	0.0000
DUMMY_DEB	-1.289683	0.091524	-14.09122	0.0000	DUMMY_NL4	-0.526179	0.092987	-5.658649	0.0000
DUMMY_DEC	-1.610840	0.143349	-11.23715	0.0000	DUMMY_PT1	0.283228	0.129076	2.194271	0.0283
DUMMY_DED	-0.532574	0.158572	-3.358559	0.0008	DUMMY_SE0	0.433952	0.120733	3.594322	0.0003
DUMMY_DEE	-1.254139	0.165024	-7.599754	0.0000	DUMMY_UKC	-0.646452	0.100920	-6.405607	0.0000
DUMMY_DEF	-0.567748	0.103488	-5.486135	0.0000	DUMMY_UKD	-0.375521	0.081895	-4.585388	0.0000
DUMMY_DEG	-0.913456	0.163892	-5.573526	0.0000	DUMMY_UKE	-0.179863	0.086833	-2.071354	0.0384
DUMMY_DK0	0.143599	0.114672	1.252256	0.2106	DUMMY_UKF	-0.301364	0.081321	-3.705835	0.0002
DUMMY_ES1	-1.098204	0.095175	-11.53880	0.0000	DUMMY_UKG	-0.350168	0.084382	-4.149817	0.0000
DUMMY_ES2	-0.651567	0.084345	-7.725049	0.0000	DUMMY_UKH	-0.069641	0.091296	-0.762802	0.4457
DUMMY_ES3	-0.015755	0.091001	-0.173126	0.8626	DUMMY_UKI	0.219808	0.081899	2.683876	0.0073
DUMMY_ES4	-1.443328	0.106175	-13.59389	0.0000	DUMMY_UKJ	0.310189	0.099724	3.110482	0.0019
DUMMY_ES5	-0.111043	0.085583	-1.297494	0.1946	DUMMY_UKK	-0.305429	0.080815	-3.779376	0.0002
DUMMY_ES6	-1.058933	0.084079	-12.59443	0.0000	DUMMY_UKL	-0.668134	0.104281	-6.407039	0.0000
DUMMY_ES7	-1.804090	0.137875	-13.08500	0.0000	DUMMY_UKM	-0.124750	0.083983	-1.485422	0.1376
DUMMY_FI1	0.153155	0.118137	1.296421	0.1950	DUMMY_UKN	-1.294967	0.123238	-10.50785	0.0000

R-squared	0.920074	Mean dependent var	3.763354
Adjusted R-squared	0.917470	S.D. dependent var	1.381296
S.E. of regression	0.396819	Akaike info criterion	1.020773
Sum squared resid	357.6033	F-statistic	353.2832
Log likelihood	-1122.367	Prob(F-statistic)	0.0000

6.4.5 Explicative Power of Direct FP Modelling

Having ensured the direct FP estimation's stability for numerous sub-samples, we now enquire the model structure's explicative power regarding its fit over the entire sample. A first look at Estimation 7 and Estimation 9 sets the R^2 for the log-transformed dependent variable at 0.853 and 0.92 for the model with and without regional dummies, respectively. For plain values, the corresponding R^2 amount to 0.801 and .0941.[126]

By construction, Estimation 7's errors are larger on the whole, but also with respect to regions: The region-specific row sums of the 68 x 68 squared residuals matrix are considerably higher than the comparable values for Estimation 9.

The scaling procedure described in section 6.1.5 adjusts the model's output for errors in estimating a region's total number of collaborative links (thus, it adjusts for mis-fitting in the first component of the matrix). This process pushes the region-specific residual sum of squares to zero and allows for concentrating on a matrix's relative goodness-of-fit. The fit of Estimation 9 is only slightly ameliorated by the scaling procedure, its R^2 amounts to 0.945. Estimation 7, in contrast, under-estimates total collaboration numbers. Applying the scaling procedure to its outcome considerably enhances R^2 from 0.801 to 0.901 (in stacked form).

[126] The R^2 for plain values refer to the models' fit retransformation with respect to untransformed cooperation values, in stacked form: i.e. only the matrices' main diagonal and lower triangular are taken into account.

7 SYNTHESIS AND INTERPRETATION

The following pages will review the results of the two modelling approaches pursued in this study against the backdrop of theoretical considerations laid out in Part I of this study: First, the performance of the two model structures will be compared for goodness of fit. Subsequently, the parameters chosen by the corresponding selection procedures will be evaluated on the ground of the hypotheses provided in sections 2 and 3. Moreover, we will investigate region-specific effects for the NUTS-1 regions under consideration. Finally, a concluding section will highlight the implications to be drawn from the empirical analysis.

7.1 Model Comparison

Two approaches were evaluated in order to determine a model structure for regional FP4 collaboration: The direct estimation of the collaboration matrix, and the approach via the estimation of implied masses and implied distances and their subsequent combination. If there is a “true” model structure dependent on the factors evaluated, and if the explorative approach followed is well designed, the two explorative approaches would have yielded exactly the same, clear-cut structure. To a considerable extent the variables included by the two “final” structures overlap: apart from *PATSTRUCT*, all of the variables included in the “direct” model also are relevant in the final “indirect” structure. However, the two models did not yield exactly similar results, as would have been the case under perfect conditions.

This leaves the question which of the two structures is the “better” approach, in particular which of both reflects reality more accurately. Statistical indicators such as R^2 or log-likelihood provide information about the models’ fit: With respect to these statistics, the indirect approach performs slightly superior; however no major differences in performance emerge.

Table 11: R^2 for direct and indirect approach results

		direct		Indirect	
		without reg. dummies	with reg. dummies	without reg. dummies	with reg. dummies
unscaled	log-transformed	0.853	0.919	0.850	0.874
unscaled	Plain values	0.801	0.941	0.850	0.888
Scaled	log-transformed	0.909	0.917	0.920	0.925
Scaled	Plain values	0.901	0.945	0.953	0.960

As is shown in Table 11, the fit of the “directly” estimated model structure (without regional dummies – Estimation 7) is inferior to that of the combined estimation results for implied indicators (as in section 6.3.2). However, adding regional dummies to the basic direct structure (Estimation 9) substantially increases performance to an R^2 considerably above that of “implied” estimation – while an addition of regional dummies to implied distance estimation affects R^2 only marginally.

The approach via implied indicators seems to perform slightly superior – in particular with respect to scaled values, which represent a fit adjusted for a large share of mass-specific effects. However, the differences in R^2 are too minor to allow for a decisive assessment.

In order to determine the “better” of the two model structures, we resort to an encompassing test, a frequently used method following simple reasoning as outlined by Greene (2003):

(38) $$y_t = (1-\gamma)\hat{y}_{t,A} + \gamma\hat{y}_{t,B} + \bar{u}_t$$

As is shown in (38), the quality of two model’s fits may be determined by regressing the actual dependent data on them. The dependent data y can be represented as a weighted average of the fit of model A $\hat{y}_{t,A}$ and fit of model B $\hat{y}_{t.B}$, plus a composite error term $\bar{u}_t$. If the weighting factor γ is either 1 or 0 then one model fit in (38) encompasses the other. Deducting $\hat{y}_{t,A}$ from both sides of (38) yields a simple structure to test for γ as in (39).

(39) $$\hat{u}_t = y_t - \hat{y}_{A,t} = \gamma\left(\hat{y}_{B,t} - \hat{y}_{A,t}\right) + \bar{u}_t$$

The structure in (39), i.e. the residuals of model A regressed on the difference between the two model’s fits, represents a encompassing test “whether model B encompasses model A”; in other words whether γ is significantly different from zero. For the “implied” and “direct” model structures presented in this study, Estimation 10 represents an encompassing test for plain values (out of estimations without regional dummies).

Estimation 10: Encompassing test whether implied model fit encompasses direct estimation fit

Dependent Variable: COOPFP4-1-COOPFP4FD
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
COOPFP4FIMP-COOPFP4FD	0.642389	0.016026	40.08382	0.0000
R-squared	0.396714	Mean dependent var		8.480306
Adjusted R-squared	0.396714	S.D. dependent var		65.76286
S.E. of regression	51.07897	Akaike info criterion		10.70505
Sum squared resid	6118249.			
Log likelihood	-12556.02			

By definition, the coefficient resulting from the reverse of Estimation 10 (i.e. a test whether the direct model fit encompasses the implied fit) is exactly one minus the coefficient in Estimation 10. This fact is demonstrated by Table 12, which displays coefficients, the corresponding t-statistics and R^2 for encompassing tests whether “the direct structure’s fit encompasses that of the implied approach”, and the (independently estimated) reverses of these encompassing tests. Moreover, Table 12 shows encompassing tests among log-transformed fits, fits in "plain" levels and scaled fits; in each case either the direct fit without regional dummies (Estimation 7) is opposed to an implied fit not deriving from regional dummies (Estimation 2 and Estimation 7), or both fits were compiled with the inclusion of regional dummies.

Table 12: Encompassing test results

	direct encompasses indirect						Indirect encompasses direct					
	without dummies			with dummies			without dummies			with dummies		
	coef	t-stat	R^2	coef	t-stat	R^2	coef	t-stat	R^2	coef	t-stat	R^2
log-transformed	0.537	15.898	0.097	0.852	37.380	0.373	0.463	13.733	0.074	0.148	6.484	0.018
Plain	0.358	22.314	0.170	0.723	51.806	0.527	0.642	40.083	0.397	0.277	19.804	0.125
scaled log-transformed	0.125	2.678	-0.02	0.281	5.892	-0.01	0.875	18.727	0.119	0.719	15.080	0.074
scaled plain	0.156	10.672	0.046	0.270	12.432	0.061	0.844	57.660	0.586	0.730	33.619	0.324

In all of the encompassing estimations, the respective coefficient standard errors are fairly low on either side – therefore γ is both significantly different from one and from zero. In most cases, the coefficient for the direct estimation is lower than the one for the implied approach: only in unscaled log-transformed representation with regional dummies, the direct approach appears to exhibit "better" performance. Neither of the two structures can be dismissed based on encompassing test results. Therefore the concluding sections will draw on the outcomes of both models.

7.2 Explanatory Factors: Review and Interpretation

Which factors are responsible for interregional FP4 collaboration? The variable selection procedure outlined in section 6.1.4 reduced the set of eligible explanatory factors to a handful of factors consistently affecting all of the regions under consideration. Due to the agglomerate nature of the data investigated, the chosen factors rather point to the environmental aspects described in section 2.2 than to the microeconomic factors listed in section 2.1.

Against the backdrop of the gravity model concept, each of the chosen factors may attributed to one of three dimensions: Interregional "distances", node-specific variables shaping "mass" and, between those two, the core-periphery dimension. The latter two dimensions were explained by n x 1 vectors defined as "plain" or as efficiency-related mass factors presented in section 5.3, while the former is both affected by core-periphery considerations and n x n "distance" matrices.

Nine factors survived the selection procedure for the "direct" estimation approach (including the core-periphery factors from Estimation 1) while ten factors were chosen for the "implied" approach. Among those factors, relative importance may be determined by standardised coefficients, the coefficients corresponding to a linear structure comprising centred rather than original data series. Centring data implies correcting original data for mean and variance in order to obtain series with zero mean and uniform variance. If a linear structure in original data is given in (40) and since arithmetic average of the dependent is equivalent to (41), the structure in centred data can be represented as in (42).

(40) $$y_j = \beta + \sum_i \alpha_i x_{i,j} + \varepsilon_j$$

(41) $$\bar{y} = \beta + \sum_i \alpha_i \bar{x}_i$$

(42) $$\frac{y_j - \bar{y}}{\sigma_y} = \sum_i \alpha_i \frac{\sigma_i}{\sigma_y}\left(\frac{x_{i,j} - \bar{x}}{\sigma_i}\right) + \frac{\varepsilon_j}{\sigma_y}$$

$$\alpha_i^* = \alpha_i \frac{\sigma_i}{\sigma_y}$$

Thus correcting the original coefficients α_i by the standard deviation of series i and the standard deviation of the dependent yields the coefficients for centred data α_i^*. Since the mean and variance of standardised series are uniform, the absolute size of standardised coefficients α_i^* represents an indicator of the relative impact for each series in a single structure. Table 13 displays the standardised coefficients for the 14 data representations in the structures for the core-periphery dimension (Estimation 1), the implied approach (Estimation 2 and Estimation 5) and the direct approach (Estimation 7).

Table 13: Standardised coefficients for selected estimation results

Refers to Dependent	Estimation 1 *L_IMPDISTSUM*	Estimation 2 *LL_IMPDIST*	Estimation 5 *L_IMPMASS*	Estimation 11 *L_COOPFP4*	Estimation 7 *L_COOPFP4*
L_RSTAFF_TOT	-0.721	-	0.808	0.232	0.169
LL_RDXPNAT_T	-	-	-	-	0.201
RDXPNAT_BIZ	-	-	0.309	0.129	-
PAT_TOPBIZ	-	-	-0.328	-0.066	-
L_IMPDISTSF	-	0.660	-	-0.587	-0.537
DIST2CORE	0.294	-	-	-	-
RDXPC_GOV	-0.240	-	-	-	-
PUBNAT_RST	-0.291	-	-	-	-
SQRT_GEODIST	-	0.169	-	-0.050	-
L_RDXDIFF	-	-	-	-	-0.108
RDXDIFF	-	0.079	-	-0.104	-
PATSTRUCT	-	-	-	-	-0.078
INTRAROMANIC	-	-0.170	-	0.170	0.156
DUMMY_DDR	-	-	-0.169	-0.166	-0.122

The fit of Estimation 1 provides the indicator *L_IMPDISTSF*, a constituent of the structures Estimation 2 and Estimation 7. While Estimation 7 directly specifies the fitted log-transformed collaborative links *L_COOPFP4*, the corresponding implied approach is a nearly-linear combination of Estimation 2 and Estimation 5. In order to render coefficients for the implied approach roughly comparable to those of Estimation 7, we introduce Estimation 11, whose standardised coefficients are also displayed in Table 13.

Estimation 11: Log-transformed collaborative links regressed on indicators selected for “implied“ approach

Dependent Variable: L_COOPFP4
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-16.98020	4.949788	-3.430491	0.0006
L_RSTAFF_TOT	0.234899	0.018051	13.01323	0.0000
RDXPNAT_BIZ	1.406688	0.092177	15.26066	0.0000
PAT_TOPBIZ	-31.46054	4.628434	-6.797232	0.0000
DUMMY_DDR	-0.657701	0.033540	-19.60922	0.0000
L_IMPDISTSF	-1.479690	0.043411	-34.08562	0.0000
RDXDIFF	-0.845998	0.073470	-11.51484	0.0000
SQRT_GEODIST	-0.005340	0.000969	-5.513089	0.0000
INTRAROMANIC	0.638887	0.032001	19.96476	0.0000

R-squared	0.850659	Mean dependent var	3.763354
Adjusted R-squared	0.850148	S.D. dependent var	1.381296
S.E. of regression	0.534710	Akaike info criterion	1.589644
Sum squared resid	668.1822	F-statistic	1663.970
Log likelihood	-1855.652	Prob(F-statistic)	0.000000

For the factors included both in Estimation 7 and in the implied approach, the standardised impact indicators α_i* are about equivalent. The indicator *L_IMPDISTSF* apparently has the greatest (negative) impact for both model structures. However, *L_IMPDISTSF*, the result of Estimation 1, is largely dependent on *L_RSTAFF_TOT*, the second-most important indicator in both structures. Multiplying the α_i* for *L_RSTAFF_TOT* in Estimation 1 with the coefficients for *L_IMPDISTSF* in the latter two structures yields a combined positive coefficient of about 0.4 in both cases. Thus total regional research staff *L_RSTAFF_TOT* is the most important indicator shaping FP4 collaborative links. The remaining indicators included both in the implied and in the direct approach exhibit α_i* between 0.1 and 0.2 in absolute size: While *INTRAROMANIC* has a positive impact, *DUMMY_DDR* and, slightly less important, *RDXDIFF* resp. *L_RDXDIFF* affect interregional collaboration negatively. The two factors describing intra-national importance *LL_RDXPNAT_T* and *RDXPNAT_BIZ* both exert positive influence on collaboration, albeit to a differing extent. The remaining indicators being part of Estimation 1 affect log-transformed collaboration equally with α_i* between 0.13 and

0.18 (in absolute size):[127] While *DIST2CORE* appears to have a negative effect, *RDXPC_GOV* and *PUBNAT_RST* compensate with positive impact of about equal size.

The impact of indicators appearing only in the “implied” or in the “direct” structure, in contrast, is considerable lower: *SQRT_GEODIST*, *PATSTRUCT* and *PAT_TOPBIZ* all affect collaboration negatively with α_i* between –0.05 and -0.08.

The mentioned coefficients partly belong to n x 1 “mass” data and n x n “distance” data. From a theoretical point of view, the may be grouped into five main categories, three of them comprising “mass” and two among them constituted of “distance” factors. The following pages will outline those categories, with reference to the hypotheses listed in Table 1.

“Plain” Mass

The factor found to be by far the most important is total regional research staff, *RSTAFF_TOT*. This conforms to Sharp’s (1998) central indicator for regional collaborative potential. Remarkably, *RSTAFF_TOT* beats all other indicators for collaborative potential such as HRST, research expenses and in particular sector-specific data. Moreover it is the only “plain” mass indicator to have remained in the final structures. This feature highlights the importance of persons rather than other structural indicators for the formation of FP projects.

Moreover *DUMMY_DDR* may be comprised under the “plain mass” category, although its importance is possibly only due to the fact that the Eastern German research sector was still beginning to integrate into Western European research schemes at the time.

Efficiency Mass Indicators

The efficiency ratios introduced in section 5.3.3 play a rather minor role: Only *PUBNAT_RST* (natural science publications per researcher) is included in both final model structures, via the indicator *L_IMPDISTSF*, i.e. as an indicator affecting core-periphery positioning. Its inclusion conforms to hypotheses 5a) and 5c), whereby “scientific excellence” promotes the potential for collaboration. In Estimation 1, regions with an over-proportional number of natural science publications per researcher locate themselves more centrally in the FP4 core-periphery spectrum: First, this is certainly due to the “scientific excellence” found in those regions. Second, the areas with large publication quotas are also mostly located in the economic cores of their respective countries – in particular in “blue banana” regions (compare Andersson/Persson (1993) and Hilpert (1992)).

---

[127] Multiply the standardised coefficients α_i* in Estimation 1 with the α_i* for *L_IMPDISTSF* in Estimation 7, respectively Estimation 11, to obtain their α_i* in the “final” structures.

The share of public research expenditure in regional total research expenditure *RDXPC_GOV* may be as well subsumed under the “efficiency” category: Its importance may be attributed to the fact that the share of public research organisations in FP4 collaboration (28%)[128] is by far higher than the share of governmental research expenditure in total EU-15 research expenditure (15%). R&D resources in the governmental research sector thus over-proportionally foster FP participation. In addition, the distribution vaguely reflects the general pattern of regional involvement in the FP: Central, less research-intensive continental regions (DE3, GR3 or ITE, for instance) exhibit higher *RDXPC_GOV* than both peripheral poor regions (e.g. ES3, ITG) and research-intensive, rich blue-banana and northern regions (e.g. SE0, DEA, UKL). The former are generally found to collaborate more within the FP4 than the latter (compare section 2.2.6 for that purpose).

The factor *PAT_TOPBIZ*, representing the number of patents per Amadeus 1,500,000 firms *TOPBIZ*, is only included in Estimation 5; its coefficient is unexpectedly negative. Interestingly, the overwhelming majority of patent-related mass indicators considered in this study feature negative coefficients (if they proved significant). Factors attributed to the Amadeus 1,500,000, in contrast, unanimously contributed positively to collaboration/implied masses. In particular, *TOPBIZ* performed considerably superior with respect to each other firm number-related indicator.

It is relatively straightforward to interpret *TOPBIZ* as a factor positively affecting mass: The literature review already hinted at the number of large firms to be an important element of absolute collaborative potential. In comparison to the total number of firms, or the number of large firms, *TOPBIZ* constitutes an even better indicator: First it is more likely to constitute a complete sample; second, it comprises “important” rather than only large firms – thus it more easily encompasses small research or consulting boutiques of small size, but of relevance to European research. This reasoning would confirm hypotheses 1a) to 1d) in their best sense. But why does the ratio of patents to *TOPBIZ* feature a negative coefficient? This fact is explicable by Estimation 5’s structure: It has already been mentioned that for the 68 NUTS-1 regions evaluated, the share of firms in FP4 projects (36%) is considerably lower than the share of business research expenditure in total (67%) or that of business research staff in total (58%), and patents primarily stem from the profit-oriented sector. It has also been mentioned that many “excellent” regions like SE0, FI1, FR1 or the large German *Länder* participate less in FP4 than their structural indicators would imply. Finally, patents are regarded as indicators of applied research, while the large share of public organisations in FP4 projects implies that the FP is more oriented towards basic or “pre-competitive”

[128] Compare Table A.3 in the appendix.

research: The picture leads to the conclusion that, adjusted for mass indicators such as *RSTAFF_TOT*, efficient applied business research does not necessarily foster a region's representation in FP4. This effect may partly be attributed to the European commission's efforts to achieve a "fair" distribution of FP funds. If hypotheses 5a), 5b) and 7a) are to be measured by patent-related indicators, then the negative sign of *PAT_TOPBIZ* represents an argument against those mechanisms to hold in the case of the FP.

Finally, we consider an efficiency indicator eventually not included in either the "direct" or the "implied" structure: regional research expenditure per researcher *XPRES_BIZ*, *XPRES_EDU, XPRES_GOV* and *XPRES_TOT*. Although expenditure per researcher, particularly in the business sector (*XPRES_BIZ*), is positively correlated with log-transformed collaborative links (or implied masses), the effect turns significantly negative when it is included in a greater structure. In particular *XPRES_GOV*, the only of those indicators not to be positively correlated with *L_COOPFP4*, exerts strong negative influence on collaboration adjusted for the highlighted indicators. This would imply that comparatively less-than-well funded public research spurs participation in the FP4, conforming to hypotheses 8b) and 8c). Although the negative effect persists for all regional sub-samples, *XPRES_GOV* was finally included in neither structure for the considerably varying size of its impact over regions.

Intra-National Cohesion Indicators

The n x 1 indicators *RDXPNAT_TOT*, *RDXPNAT_BIZ* and *DIST2CORE* reflect regional core-periphery location on the national level. The factors *RDXPNAT_TOT* and *RDXPNAT_BIZ* reflect the share of regional research expenditure (overall, respectively in the business sector) in the national total, both featuring a positive effect on total collaboration or implied mass. *DIST2CORE*, the geographic distance to the national "economic core", contributes to *IMPDISTSUM*, the proxy for peripheral position in the relationships among regional nodes.

Remarkably, this effect is barely considered in academic literature on FP collaboration. Only Sharp (1998) was to investigate the regional dimension and to highlight the suspicion that economic core regions in cohesion countries may be favoured over the corresponding national periphery. While Sharp (1998) eventually did not find evidence to support the hypothesis, the importance of the cited factors clearly states that regions considered as central are more involved in FP collaboration than could be expected from structural indicators.[129]

---

[129] This relationship also holds when the 100%-observations DK0, FI1, IE0, LU0, PT1 and SE0 are excluded from the sample.

The importance of those factors could be attributed to genuine efficiency factors; geographic concentration may advance research efficiency through network economies in an integrated local research "market". However, such efficiency-through-scale factors could have been easily captured by the numerous other efficiency variables in the sample. The significance of national-level importance must thus be attributed to other effects. Two main effects can be imagined: First, there may be a clustering of top-level institutions at the national level – in most large corporate groups, for instance, there is a single national headquarter in each country of operation – regardless of the country's size. Similar features can be reckoned for national champions in public research. And according to anecdotic evidence, those top-level institutions tend to be concentrated in economic core regions, such as Lombardy, Paris, or Randstad. Other research employment, or specialised consultancies may be equally found in such central areas – akin to the concept of a technological district. Moreover institutions in these central regions may dispose of higher international reputation and easier access to prospective foreign collaborators than similar regions in other nations exposed to stronger national competition.

Second, the cohesion-related efforts of the European Commission may either cause the observed pattern or at least aggravate national-level clustering as mentioned before. As was outlined in section 3.1.6, the European Commission cares for an "equal" allocation of funds to member states, even if this aim is not explicitly stated in the FP objectives. On these grounds, weaker member states such as Greece or Portugal are perceived to play a larger role in the FP than their number of researchers, research expenditure, etc. would imply (Sharp 1998). However, the Commission apparently puts less emphasis on an equal distribution of funds over regions. I.e. it cares for the inclusion of Spain, for instance, no matter if the participating organisation is located in Barcelona or in Seville. While the fair distribution of FP projects among EU-15 member states aims at reducing disparities, it does not address intra-national differences and even allocates peripheral regions less participation than their structural indicators would imply.

Heterogeneity as a Distance Factor

The final model structures include two $n \times n$ "distance" indicators measuring homogeneity versus heterogeneity, namely *RDXDIFF* and *PATSTRUCT*. The latter represents dissimilarity in patent applications per broad economic sector. The former factor measures relative dissimilarity in the importance of the governmental, educational and corporate research sectors: The factor's negative coefficients imply that increased disparity in the importance of sectors in two regions' research leads to decreased bilateral collaboration. The significance of *RDXDIFF* is broadly in line with hypotheses 15a) to 15d) which highlight the impediments to public-private research collaboration, but also with 25) which reckons the prospects for spillovers to correspond to technological similarity. Hypothesis 25) is also expressed in the impact of *PATSTRUCT*, since Hussler (2003) uses a very similar approach in testing for this impact. To some extent, the patents- and thus business-related factor *PATSTRUCT* supports hypothesis 12b), which sees firm collaboration fostered through symmetry in production and research.

Even if these hypotheses cannot be extrapolated to the agglomerate level, different sector orientation between regions clearly seems to hinder bilateral cooperation within the FP.

Cultural Affinity as a Distance Factor

The initially introduced n x n culture and language distance indicators did not weather the variable selection procedure. Hofstede's (1980) cultural dimensions fared badly, though this could be attributed to their inherent limitations frequently criticised in related literature. Language distance indicators performed slightly better (with expected signs) but were eventually also dropped from final model structures due to their differing impact across regions.

The dummy variable *INTRAROMANIC* in contrast proved highly significant and stable. The collaborative links between two regions are raised by about 17%-18% if both interacting regions belong to a Romanic-language country (or Greece).[130] This effect was not to be expected from the literature review but was introduced later on when exploring the patterns of implied distances. In addition, Table 6 shows that Romanic countries are considerably more linked with France than other member states.

Strikingly, the indicator beat all other affinity factors such as *CULTDIM* or *LANGSTUD*; an opposing *INTRAGERMANIC* dummy did not show any impact at all. As with most dummy factors, the importance of *INTRAROMANIC* may just reflect the impact of factors not observed or not included into the sample of eligible explanatory variables. There may be a finer-than-measured effect of cultural affinity or common language, in line with hypotheses 22) and 23). Almost certainly, the prospect for personal contact among researchers from Romanic countries is higher than for the European average. It is questionable though whether those factors provide the only grounds for the intra-Romanic effect observed.

---

[130] If Estimation 7 or Estimation 11 are retransformed into plain values, their additive structures become multiplicative: The dummy factor *INTRAROMANIC* then does not matter if *INTRAROMANIC* is zero (the remaining factors and coefficients is multiplied by exp(0)). If both nodes involved into a link are from the Romanic area, the result from the remaining terms are multiplied by exp(0.156) = 1.168 or exp(0.17) = 1.185.

Geographic distance *GEODIST* is as well categorised under “cultural affinity” since it is viewed as a proxy for personal contact prospects based on factors exogenous to the research sector. It finally only made it into the implied distance model Estimation 2, not into Estimation 7 obtained through the “direct” approach. For the reasons outlined in section 5.2.1, geography seemingly matters, but to a considerably smaller extent than in the economics of international trade (compare Porojan, 2000). Other distance factors, especially *INTRAROMANIC* and *RDXDIFF,* seem to be of greater importance.

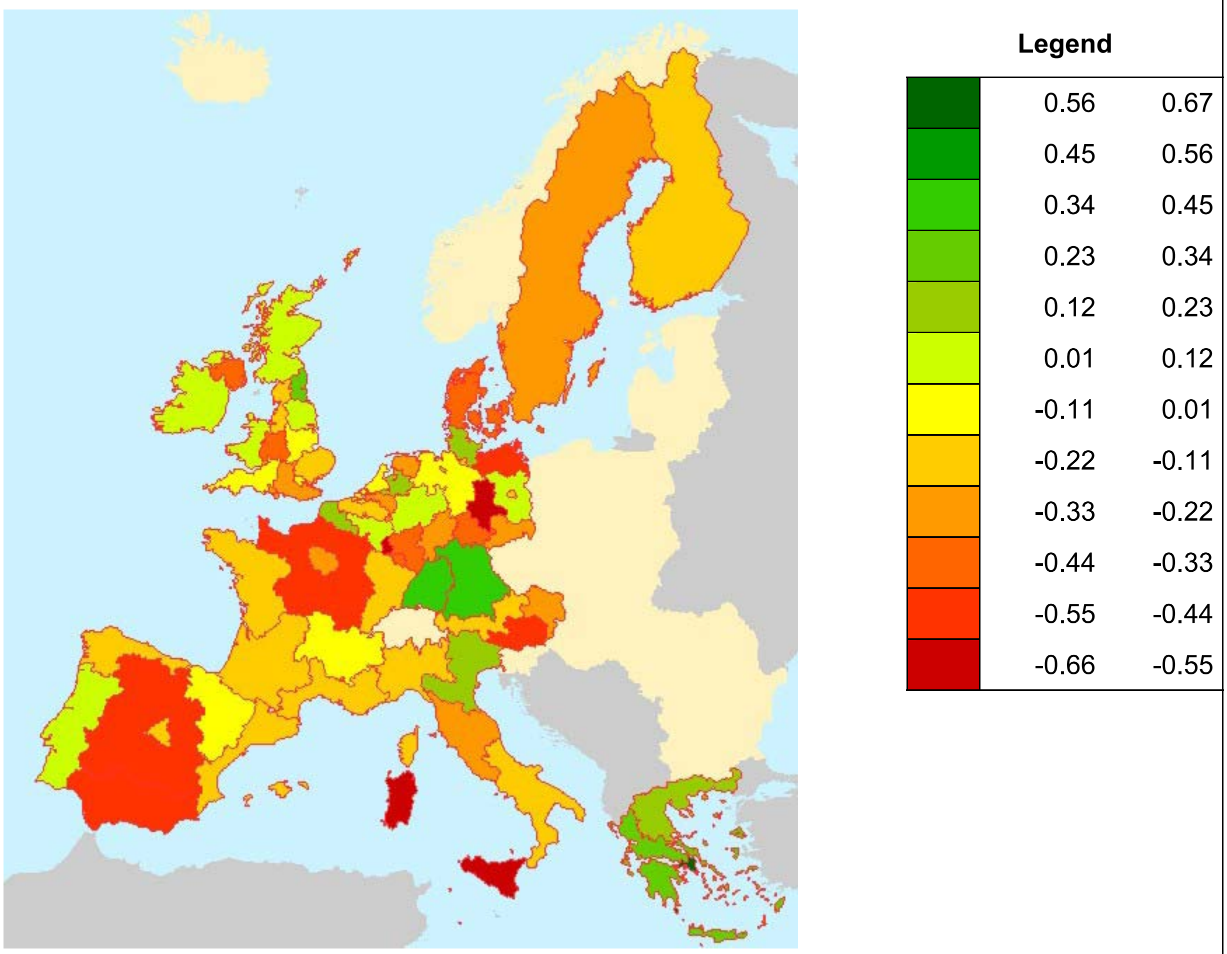
7.3 Node-Specific Effects

Given the general impact of mass and distance factors, regional dummies and residuals hold additional insight: With respect to where the model structures place NUTS-1 regions, they represent the extent to which a region is over- or under-proportionally involved in FP links. We concentrate on the arithmetic mean of residuals per region – which are equivalent to a estimating regional dummies. Average residuals broadly coincide both for the implied and the directly estimated model structure – though in several cases they differ considerably. The following maps depict the sign and size of regional residual averages, where dark green (positive residual) signifies under-estimation (i.e. a region is actually more involved into research collaboration than expected from a model) and dark red symbolises over-estimation (i.e. a region collaborating less then predicted).

Differing signs and magnitude may be attributed to two potential causes: First, the differing model factors and coefficients from the two approaches could result into adversary outcomes: This is the most likely for regions whose residual averages are relatively far apart in the two model approaches concerned – for instance the Sweden and Denmark, or the Southern Netherlands.[131] Second, both models could match regional collaboration intensity accurately enough, such that regional average residuals are purely erroneous.

[131] Remarkably, Sweden and Denmark exhibit negative residual averages in the implied estimation approach – even if the implied distance estimation includes geographic distance (tearing down estimates for “remote” regions), while the direct approach does not.

Figure 12: Regional average residuals from implied approach (section 6.3.2)[132]



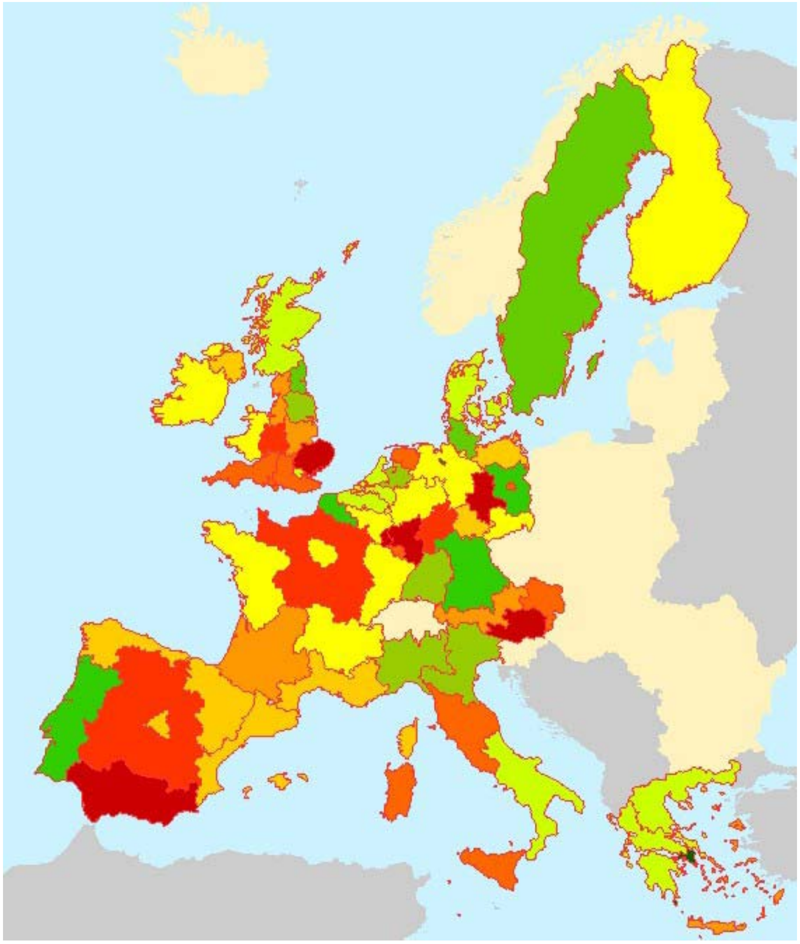
However, both explanations do not accurately match reality: Regional residual means are far too high compared to their expected value of zero in order to be attributed to pure erroneous factors. Moreover, those regional residuals means are too related to be only dependent of model specificities. At least for the regions exhibiting residuals of comparatively equal magnitudes over both regions, regional fixed effects are assumed to play a role.

The following paragraphs will highlight those fixed effects and their consistency over models by country: A quick glance at Figure 12 and Figure 13 reveals that at the country level positive and negative residual averages are distributed fairly evenly among EU-15 member states.

---

[132] Regional row sums over residuals from log-transformed implied estimation results: *L_COOPFP4* minus results from Estimation 5 combined with results from Estimation 2.

Figure 13: Regional average residuals from direct approach (Estimation 7)[133]



Legend

0.42	0.505
0.334	0.42
0.249	0.334
0.164	0.249
0.078	0.164
-0.007	0.078
-0.093	-0.007
-0.178	-0.093
-0.264	-0.178
-0.349	-0.264
-0.434	-0.349
-0.52	-0.434

Nearly the total of French and Spanish regions appear with negative residuals, i.e. are over-estimated by both modelling approaches: In ***France***, only FR3 exhibits relatively large, positive residuals (i.e. it collaborates more than implied by the models), while the rest displays moderately negative numbers. While FR7 performs a bit better than its national counterparts, FR2 appears as the worst-performing in relation to its structural potential as estimated by the models. This may be attributed to the fact that FR2 is entirely focused on Paris and lacking a centre of its own. Moreover its research capacity could be attributed to outlets headquartered in Paris, with corresponding collaborative links consequently accounted for by the capital (FR1). Notably, FR1 itself exhibits moderately negative to zero residuals, although it is the most "collaborative" region in the entire FP4 spectrum.

[133] Regional row sums over residuals from log-transformed direct estimation results: *L_COOPFP4* minus results from Estimation 7.

Spanish regions appear as well to be consistently over-estimated by both model structures. Apart from ES7 (not depicted) all regions' residual averages bear negative signs. Madrid (ES3), the North (ES1 and ES2) and East (ES5) cooperate only slightly less than they should according to the models, whilst the poorer Centre (ES4) and South (ES6) are hugely over-estimated. The poor performance of ES4 could be attributed to similar factors as for the French Centre FR2. ES6, in contrast, is a large region with important urban centres – its poor performance is at first sight startling particularly when compared e.g. to the peripheral Greece or ITF.

Among the small cohesion countries Greece, Ireland and Portugal being under-estimated in their collaboration intensity, ***Greece*** stands out with the highest residual averages in the EU-15 area. Particularly its core region Attica (GR3) exhibits by far the largest values in both the direct and the implied models. However, the frequently marginalized regions GR1 and GR2 evenly perform over-proportionally well, given model expectation. The ambiguous performance of the island region GR4 can be attributed to its relative smallness in size. The intense Greek involvement into the FP programmes has already been noticed (Geuna 1998, p. 684; Hernán/Marin/Siotis 2003, p. 87), and Figure 13 once again confirms Greek over-performance. Both modelling approaches failed to explain the reasons for the strong Greek standing, even if this was known beforehand. The reasons for Greek, and particularly Athens' collaboration intensity and its potential benefits for the country could constitute an interesting matter for future research. (Hakala/Kutinlahti/Kaukonen (2002) hint at a possible reason, namely that the inavailability of funding on the national level may spur organisations to apply for supra-national funding).

Similar to Greece, ***Portugal*** (PT1) exhibits zero to positive residual averages. The according intra-national distribution however cannot be determined. ***Ireland*** (IE0) appears as a similar case, ranking slightly below Portugal in both Figure 12 and Figure 13 .

The collaboration intensity of Northern ***Italian*** regions ITC and particularly ITD seem under-estimated by both modelling approaches, while the Southern ITF appears with residual means of about zero. The central ITE and insular ITG, in contrast, exhibit considerably negative residuals, with the latter being over-estimated to a similar extent as ES6. The Italian North-South division appears familiar, with a cooperative, innovative industrial North and a structurally weak South: In particular the island regions are known for not drawing on the full potential of funds available from European sources. The continental south ITF, in contrast, appears to be moderately well involved in FP collaboration with respect to its research resources. Interestingly the research expenditure of underperforming ITE is marked by a record share of governmental research (40%),[134] far higher than for its northern and southern counterparts. The picture fits well into complaints about an inflated state sector in Rome, but would have to be examined more closely in order to derive decisive statements.

Both models over-estimate ***Austrian*** regions to an about similar degree, with the Western (AT3) region, followed by the Eastern region (AT1), exhibiting already considerably large negative residuals, while the poorer, peripheral South (AT2) ranks among the worst performing regions in the EU-15.

The ***German*** *Länder* are characterised by a comparatively heterogeneous pattern: Among the large regions, the southern and economically advanced regions of Bavaria (DE2) and Baden-Württemberg (DE1) figure among the most under-estimated areas in the EU-15 (along with smaller Schleswig-Holstein (DEF) and Bremen (DE5)). North-Rhine-Westphalia (DEA), the largest region, exhibits residuals only marginally above zero, with Lower Saxony (DE9) and Hamburg (DE6) ranking slightly inferior. The large *Land* of Hesse (DE7), however, is fairly over-estimated in its collaborative activity, accompanied by its neighbours Rhineland-Palatinate (DEB) and Saarland (DEC).

The Eastern *Länder* of Germany exhibit particularly low numbers: The results shown in the figures do not account for the negative effect of the *DUMMY_DDR* dummy, a fixed effect for those regions (ex-Berlin). In both Estimation 5 and Estimation 7, the coefficient for *DUMMY_DDR* is about -0.4. Hence Eastern German regions are considerably more under-estimated than indicated by Figure 12 and Figure 13; in fact they rank among the regions with fixed effects the farthest below zero. As has already been mentioned, we attribute this effect to the transformation, which then was only in the beginning for Eastern Germany.

On balance, an overall weakness of German involvement into the FP cannot be identified. Instead, German collaboration within the FP is geographically distributed in an extremely heterogeneous fashion. Moreover, not only the structurally weak, but also some relatively well-off regions such as DE7 or DEB are considerably over-estimated.

All of the three ***Belgian*** regions display zero to moderately positive residual averages for both approaches evaluated. This implies that neither the positive fixed effects as mentioned by Geuna (1998), nor considerable differences between the Flemish and the Walloon regions were found. ***Luxembourg*** ranks among the worst-performing NUTS-1 regions given its resources – however this peculiar standing can be attributed to its unique characteristics as a EU-15 member state with no renowned universities and the importance of business sectors unlikely to conduct research in the FP context (such as finance).

[134] Data Source: Eurostat (2003)

Concerning ***Dutch*** regions, the comparatively weaker NL1 is consistently over- while the eastern NL2 is consistently under-estimated. The Southern NL4 shows ambiguous results. It has to be noted that NL3, the economic core region, displays values ranging from zero to moderately positive. Thus an overall “over-performance” of the Netherlands (As reckoned in Geuna 1998) cannot be confirmed.

The residual averages for ***United Kingdom*** regions display a relatively uniform picture across both models. In general, Scotland (UKM) and Wales (UKL) are moderately under-estimated, while Northern Ireland is consistently over-estimated (UKN). Among English regions, the North East performs better than expected, London about as estimated by the models, while the remaining regions cooperate less intensely than according to expectation. Interestingly, the structurally weak UKC fares considerably better than expect, along with UKE. Moreover, the “Oxbridge” UKH is over-estimated, particularly by the direct estimation approach.[135] The remaining Southern English regions are moderately over-estimated – similar to other regions surrounding large agglomerations (such as FR2, ES4, or DE4 correct for *DUMMY_DDR*). Overall, English collaboration seems to be concentrated in London which performs slightly superior to Paris, with the academic strongholds of Scotland and Wales playing their part. While “Oxbridge” appears not to involved in the FP4 as it ought to according to the models, the North-Eastern post-industrial regions set them apart from their similarly structured peers in the Midlands.

Finally, the ***Nordic*** countries seem prone to inconsistencies in the models: Sweden and Denmark immediately strike the eye as being considerably over-estimated in the implied approach but under-estimated in the direct approach. To a lesser extent this impression comprises Finland as well. The difference is the more startling, as implied approach estimation results are negatively affected by geographic distance (which has the strongest impact in the geographically remote Nordics), while it is not included in the direct approach. The exogenous factors particular to either structure do no account alone for this difference. Rather it can be attributed to the differing coefficients for *RSTAFF_TOT* (compare Table 13). Overall, no conclusion can be drawn on over- or underperformance for Nordic countries in the FP4 with regard to structural factors; the results even question the validity of statements on other NUTS-1 regions.

---

[135] The reasons for this difference are not easily identifiable, since the factors not included in both models do not show particularly extreme effects in either direction – evaluating single values, we attribute this fact to a combination of *PAT_TOPBIZ* tearing down the implied estimation result and *RDXPNAT_TOT* having as stronger in impact in the direct approach than does *RDXPNAT_BIZ* in the implied approach.

7.4 Conclusion and Implications

The 169 pages of diploma thesis up to now prepare the remaining four, which are dedicated to the broad conclusions from the paper. After a review of this study's methods, the effects behind explanatory factors are summarised, and a common regional distribution and disparity is sketched. Suggestions for further research conclude the section and, fortunately, this thesis.

Scope of This Study

The gravity model approach pursued in this study certainly holds explicative value for the macro-data constituting the basis of the paper: The impeding and promoting factors divided between node-specific "masses" and "distances" characterising inter-nodal relationship both proved valuable in the analysis. Transformation of collaboration data into the two representations provides interesting opportunities for interpretation. Moreover, the "implied" estimation based on this separation delivered results slightly superior to a "direct" estimation of regional collaboration.

Lacking a theoretic literature base on research collaboration from a macro perspective, this thesis concentrated on explorative analysis of collaboration data and prospective explanatory factors, with the latter drawing on general papers on research cooperation.

The choice of explorative analysis followed the approach primarily adopted by economic literature on public or public-private R&D cooperation (compare section 2.1). The mentioned publications concentrate on evaluating likely exogenous variables via significance tests in a single-model, whole-sample framework, with rather small data samples. In contrast, the vast amount of collaboration data available for this study allowed for introducing additional cross-check procedures in order to ensure validity and consistency of the final model structure for all 68 nodes evaluated. Furthermore, two different dependent variable formulations allowed for two separate factor-choice procedures starting from scratch – which eventually singled out an about similar set of explanatory factors, with similar impact on collaboration in both models.

Nevertheless, the results of predominantly explorative analysis have to be treated with prudence. Results in detail, especially when specific to a region or being focused at the precise effect of a variable, may be subject to debate. From an overall perspective, however, at least several broad conclusions can be drawn (to be outlined subsequently). Their main points differ from predominantly micro-level economic research on R&D collaboration: In fact, several of the major conclusions and policy implications from this study are known to have been addressed only by Sharp (1998) – to our knowledge also the only author to investigate macro-level FP data on a regional basis.

Broad outline of relevant mechanisms

Concerning further work on the FP's regional dimension, but also relating to policy and design issues of the programme, several broad implications can be deduced from the analysis conducted in this study:

- Three categories of factors were evaluated: research staff, research expenditure (each divided into three sectors) and the Eurostat-defined human resources in science & technology (HRST). As may have been expected, total research staff was the single most important factor in the final model structures. This points to the personal effort rather than capital resources as being the major driver for initiating research collaboration in the FP context.
- Similarity in sector structure promotes bilateral agglomerate collaboration. This relates both to similarity in the broad research sector structure, as to similarity in the more narrowly defined patent output structure.
- The importance of both research staff and sector similarity favour the notion of absorptive capacity as a promoter of spill-over flows with its consequences for formal collaboration: The former factor increases the likelihood of encountering a specialist for a certain sub-discipline in a region. The latter increases the potential for interaction and mutual understanding, implicitly drawing on the assumption that communication norms enable communication primarily within a scientific discipline.
- Most of the factors crucial to micro-oriented literature on FP collaboration, such as firm size, etc. were not found to have a major impact on the macro-level.
- As a striking cultural feature, Romanic or Greek-speaking regions are more likely to interact with each other than with other regions. Language or cultural factors do not explain this effect; moreover the corresponding factor proved consistently well in variable evaluation procedures. The reasons for this effect almost certainly stem from cultural affiliation. However, it could be thought over discouraging this tendency, more so since it apparently reduced the pool for collaborating with the most innovative regions for no "hard" reason. Remarkably no such effect exists for Germanic regions. In addition, geographic distance may affect collaboration patterns – if it does, however, it is confined to a minor role.
- Efficiency factors exert only minor impact on a collaborating region's "mass" and are seldom consistently included in both model structures. Most notably, publications per researcher (promoting) and patents per top firm (impeding) are included among efficiency factors: Their signs reflect the structure of FP as being far more oriented towards public research than implied by structural indicators (private-firm participation

in FP is fairly small compared to its importance to research in general). In a similar fashion, publications frequently are attributed to basic research, while patents are view as indicating applied research. Both factors support the view that the FP is far more basic-research oriented than implied by its statutes.

The Regional Dimension of FP4 Collaboration

Several conclusions can be drawn from regional fixed effects: On the country level, both model structures over- and under-estimate regions with no distinction between “cohesion” countries and member states of higher income. At the regional level, in contrast, huge disparities are revealed. In particular, those disparities remind an effect reckoned, but not confirmed by Sharp (1998): While the European Commission tends to favour “cohesion” country participants in the FP in order to ensure an “equal” distribution of funds and projects among member stats, less effort is apparently applied to aid cohesion efforts at the intra-national level.

Both models count the research expenditure of a region as a share of the national total among their most important explanatory factors (with positive impact). I.e. after accounting for plain mass factors such as research staff, the more important a region is in the national (not the European!) research system, the more it attracts research collaboration. This effect assigns national core regions an over-proportional share of collaborative links. The importance in the European spectre thus plays minor role – it is its national importance that raises a region’s collaborative “mass”. Conversely, the more peripheral and/or the smaller a region is, the less collaborative links it exhibits for given research resources.

Residual analysis shows this effect to be aggravated for countries characterised by major core-periphery disparities – such as England, Spain, France, Greece, Italy and Austria. Remarkably, many of those peripheral regions are large enough to potentially sustain scientific excellence on their own, disposing of significant agglomerations and universities. Nevertheless, those regions perform worse than their more advanced domestic counterparts, even if the models account for their less elaborated resources and for their minor importance at the national level. In particular this concerns the central and southern regions of Spain and Italy, Southern Austria, Southern England and Midlands and the French South-West. In addition, Eastern German regions are significantly less involved in the FP4 than the entirety of EU-15 regions. We attribute this effect at least partly to the restructuring process during the mid-1990s.

While the peripheral regions underperform their countries’ cores, the absolute level of residual analysis reveals sharp differences: For instance, peripheral Greek regions are consistently under-estimated by both models, while poor Sicily and Andalusia are consistently over-estimated – even if the latter comprise far more significant agglomerations and population sizes. The small cohesion countries Greece and Portugal do better than large Spain – a fact that may be attributed to the former's higher dependence on external cooperation. Moreover, the comparatively poor capital region of Athens collaborates by far more than expected by both models, while comparatively richer Rome and Madrid are over-estimated.[136] Also puzzling are the disparities among German regions, with rich Hesse and academic Rhineland-Palatinate being widely over-estimated while Bavaria as well as Baden-Württemberg are thoroughly under-estimated.

Another interesting feature: regions surrounding large metropolitan areas (central Spain, central France, Brandenburg, South-East England) collaborate considerably less than expected, while the large conurbations themselves (Paris, London, the Ruhr area, Madrid, Berlin, Randstad) are neither under- nor over-estimated.

On the hand, if cohesion and knowledge transfer to peripheral regions was an objective of the FP4, it can be deemed to have failed on the issue. On the other hand it may be argued that countries with lower resources may concentrate on creating national excellence centre at the first hand, to radiate out at a later stage. This leaves the question, however, why some regions such as the English North-East, the French North, German Schleswig-Holstein, or the Italian South around Naples perform consistently better than equally disadvantaged national or European counterparts.

Suggestions for Further Research

Sharp (1998) is the only paper known to have investigated regional involvement in the FP at the macro level. Still, both the study of Sharp (1998) and this thesis are limited in their empirical scope. Regional-level collaboration data may hold many more implications particularly relevant for public policy. On the downside, the macro-approach reveals less about individual incentives for research collaboration and its consequences to a collaborating organisation.

Further empirical research into ARCS (2003) FP data may advance into several directions. One of the less costly actions would consist of cross-checking the results of the present FP4 evaluation on FP5 data and enhance the present model structure. Evenly, consistency-check procedures could be applied to analytically more advanced methods as compared to Least

---

[136] The underperformance of Central Italy (Rome) and Madrid can be attributed to their regions comprising more hinterland, and to the fact that both capital regions do not represent the "economic cores" of their countries.

Squares: for instance, count ML (on the macro-level) or probit (on the micro-level) methods would be more appropriate given the digital nature of the data set.[137]

Moreover, the comparative analysis of disaggregated sector data, such as collaboration per key action or per organisation type would certainly provide interesting insights into the structure and specifics of the FP.

A rarely pursued, but promising approach would be the introduction of dynamics in the model framework: The impact of research collaboration over time has only been examined on the micro level in a two-stage framework evaluating spillover and cost-sharing effects on the firm level (compare Navaretti et al. 2002; Röller/Tombak/Siebert 1997). From a public policy viewpoint, the impact of alleged spillovers would be a central topic of interest regarding the financial support of technology and innovation schemes such as the FP. By connecting macro innovation data such as patents or citations with FP and national subsidy data over time, such an impact analysis might become feasible.

Considerably more data-work is involved in the gathering of explanatory data on the micro-level, directly attributable to the organisations comprised in the ARCS (2003) database. Such a data set would allow for analysis on a micro-level basis (as is the standard in literature), albeit at a much more detailed basis, on a much larger scale than ever examined; and offering the unique possibility of analysing regional data.

Stefan Zeugner, Bochum 2005

[137] Nevertheless we reckon the results from the application of such techniques to be rather similar to the results of this study.

REFERENCES

Academic Resources

Alexander A. and J.B. Ramsey (1984): The Econometric Approach to Business-Cycle Analysis Reconsidered; *Journal of Macroeconomics* 6, pp. 347–356

Andersson A. and O. Persson (1993): Networking Scientists; *Annals of Regional Science* 27, pp. 11-21.

Arrow K.J. (1962): Economic Welfare and the Allocation of Resources for Invention; in: Nelson R. (ed.): *The Rate and Direction of Inventive Activity: Economic and Social Factors*; Princeton University Press: Princeton, MA, pp. 609-625.

Baldwin W.L. and A.N. Link (1998): Universities as Research Joint Venture Partners: Does Size of the Venture Matter?; *International Journal of Technology Management* 15(8), pp. 895-913.

Banchoff T. (2002): *The Politics of the European Research Area*; ACES Working Paper: Georgetown University: Washington, DC.

Barney J. (1991): Firm Resources and Sustained Competitive Advantage; *Journal of Management Studies* 17(1), pp. 99-120.

Beckmann M. (1993): Knowledge Networks: The Case of Scientific Interaction at a Distance; *Annals of Regional Science* 27, pp. 5-9.

Button K., P. Brown, M.M. Fischer, R. Maggi, H. Ouwersloot, C. Rammer, P. Rietveld and I. Salomon (1993): *Academic Links and Communications*; Avebury Ashgate: Aldershot.

Caloghirou Y., A. Tsakanikas and N.S. Vonortas (2001): University-Industry Cooperation in the Context of the European Framework-Programmes; *Journal of Technology Transfer* 26(1), pp. 153-161.

Caloghirou Y., N. Constantelou and N.S. Vonortas (2001): *Innovation-related knowledge flows in European industry: Extent, Mechanisms, Implications – Final Report*; Final Report to the Commission, DG XII, TSER Programme, Contract No. CT 98-1118, National Technical University of Athens, Laboratory of Industrial and Energy Economics: Athina.

Caloghirou Y., N.S. Vonortas and S. Ioannides (2002): Science and Technology Policies towards Research Joint Ventures; *Science and Public Policy* 29(2), pp. 82-94.

Caloghirou Y., S. Ioannides and N.S. Vonortas (2003): Research Joint Ventures; *Journal of Economic Surveys* 17(4), pp. 541-570.

Coe D. and E. Helpman (1995): International R&D Spillovers; *European Economic Review* 39(5), pp. 859-887.

Cohen W.M., R. Florida, L. Randazzese and J. Walsh (1997): Industry and the Academy: Uneasy Partners in the Case of Technological Advance; In: Noll R. (Ed.): *Challenge to the University*; Brookings Institution Press: Washington, DC, pp. 171-199.

Cornet M. (1999): R&D Policy of the European Union; *CpB Report* 99/1, Central Plaanbureau: Den Haag.

D'Aspremont C. and A. Jacquemin (1988): Cooperative and Noncooperative R&D in Duopoly with Spillovers; *American Economic Review* 78(5), pp. 1133-1137.

Eaton J., E. Gutierrez and S. Kortum (1998): European Technology Policy; *NBER Working Paper* 6827.

Fischer M.M., T. Scherngell and E. Jansenberger (2004): *Knowledge Spillovers in High-Tech Industries across Europe: Evidence from Patent Citations*; forthcoming Working Paper, Dept. of Economic Geography & Geoinformatics, Vienna University of Economics & B.A.: Wien.

Fontana R., A. Geuna and M. Matt (2003): *Firm Size and Openness: The Driving Forces of University-Industry Collaboration*; SPRU Electronic Working Paper Series No. 103. Science and Technology Policy Research, The Freeman Centre, University of Sussex: Brighton.

Funk M. (2002): Basic Research and International Spillovers; *International Review of Applied Economics* 16(2), pp. 217-226.

Georghiou L. (2001): Evolving frameworks for European collaboration in research and technology; *Research Policy* 30(6), pp. 891-903.

Geuna A. (1998): Determinants of University Participation in EU-Funded R & D Cooperative Projects; *Research Policy* 26(6), pp. 677-687.

Greene W.H. (2003): *Econometric Analysis*, International Edition; 5th edition, Pearson Education: Upper Saddle River, NJ.

Hagedoorn J., A.N. Link and N.S. Vonortas (2000): Research Partnerships; *Research Policy* 29(4-5), pp. 567-586.

Hakala J., P. Kutinlahti and E. Kaukonen (2002): Becoming International, Becoming European: EU Research Collaboration at Finnish Universities; *Innovation* 15(4), pp. 357-379.

Hall B., A.N. Link and J. Scott (2001): Barriers Inhibiting Industry from Partnering with Universities: Evidence from the Advanced Technology Program; *Journal of Technology Transfer* 26(1), pp. 87-98.

Hauknes J. and L. Nordgren (1999): *Economic rationales of government involvement in innovation and the supply of innovation-related services*; STEP report R-08, STEP Group: Oslo.

Haynes K.E. and A.S. Fotheringham (1988): *Gravity and Spatial Interaction Models*; 3rd edition, Sage Publications: Beverly Hills, CA.

Hernán R., P.L. Marín and G. Siotis (2003): An Empirical Evaluation of the Determinants of Research Joint Venture Formation; *The Journal of Industrial Economics* 51(1), pp. 75-89.

Hilpert U. (1992, ed.): *Archipelago Europe – Islands of Innovation, Synthesis Report*; Commission of the European Union, Monitor-FAST-Program 18.

Hofstede G. (1980): *Culture's Consequences: International Differences in Work Related Values*; Sage Publications: Beverly Hills, CA.

Hussler C. (2003): Culture and Knowledge Spillovers In Europe: New Perspectives For Innovation and Convergence Policies?; *Economics of Innovation and New Technology* 13(6), pp. 523-541.

Kaiser U. (2002): An empirical test of models explaining research expenditures and research cooperation: evidence for the German service sector; *International Journal of Industrial Organisation* 20(6), pp. 747-777.

Kamien M.I., E. Muller and I. Zang (1992): Research Joint Ventures and R&D Cartels; *American Economic Review* 82(5), pp. 1293-1306.

Katsoulacos Y. and D. Ulph (1998): Endogenous Spillovers and the Performance of Research Joint Ventures; *Journal of Industrial Economics* 46(3), pp. 333-357.

Katz M. (1986): An Analysis of Cooperative Research and Development; *Rand Journal of Economics* 17(4), pp. 527-543.

Link A.N. (1981): Basic Research and Productivity Increase in Manufacturing: Additional Evidence; *American Economic Review* 71(5), pp. 1111-1112.

Link A.N. and L.L. Bauer (1989): *Cooperative Research in U.S. Manufacturing: Assessing Policy Initiatives and Corporate Strategies*, Lexington Books: Lexington, MA.

Link A.N., D. Paton and D.S. Siegel (2002): An Analysis of Policy Initiatives to Promote Strategic Research Partnerships. *Research Policy* 31(8-9), pp. 1459-1466.

Link A.N., D. Paton and D.S. Siegel (2001): *An Econometric Analysis of Trends in Research Joint Venture Activity*; Research Paper 2001/19, University of Nottingham: Nottingham.

Luukkonen T. (2002): Technology and market orientation in company participation in the EU framework programme; *Research Policy* 31(3), pp. 437-455.

Mansfield E. (1985): How Rapidly does New Industrial Technology Leak Out?; *Journal of Industrial Economics* 34(1), pp. 217-223.

Mardia K.V., J.T. Kent and J.M. Bibby (1995): *Multivariate Analysis*; 10th edition, Academic Press Limited: London.

Marín P.L. and G. Siotis (2002): *Public Policies Towards Research Joint Venture Formation: Designs and Outcomes*; CEPR Discussion Paper in Industrial Organization 3772.

Martin S. (1994): Private and Social Incentives to Form R&D Joint Ventures; *Review of Industrial Organization* 9, pp. 157-171.

Mélitz J. (2002): *Language and Foreign Trade*; CEPR Discussion Paper 3590.

Miotti L. and F. Sachwald (2003): Co-operative R&D: why and with whom? An Integrated Framework of Analysis; *Research Policy* 32(8), pp. 1481-1499.

Navaretti G.B., P. Bussoli, D. Ulph and G. von Graevenitz (2002): *Information Sharing, Research Coordination and Membership of Research Joint Ventures*; CEPR Discussion Paper in Industrial Organization 3134.

Nelson R.R. (1959): The Simple Economics of Basic Scientific Research; *Journal of Political Economy* 67, pp. 297-306.

Pohoryles R. (2002): The Making of the European Research Area – a View from Research Networks; *Innovation* 15(1), pp. 325-340.

Porojan A. (2000): *Trade Flows and Spatial Effects: The Gravity Model Revisited*; Working Paper, University of Derby: Derby.

Porter M.E. (1985): *Competitive Advantage: Creating and Sustaining Superior Performance*; The Free Press: New York, NY.

Poyago J., J. Beath and D.S. Siegel (2002): *Universities and Fundamental Research: Reflections on the Growth of University-Industry Partnership*; St. Salvator's College Economics Discussion Paper 0201, University of St. Andrews: St. Andrews.

Prahalad C.K. and G. Hamel (1990): The Core Competence of the Corporation; *Harvard Business Review* 68(3), pp. 79-91.

Pyka A. and P. Windrum (2001): *The Self-Organisation of Strategic Alliances*; Discussion Paper 209, Department of Economics, University of Augsburg: Augsburg.

Rauch J. (1999): Networks versus markets in international trade; *Journal of International Economics* 48(1), pp. 7-35.

Roediger-Schluga T. and B. Dachs (2003): *Does Technology Affect Network Structure? – A Quantitative Analysis Of Collaborative Research Projects In Two Specific EU Programmes*; Dept. of Technology Policy, ARC systems research GmbH (mimeo): Wien.

Röller L.H., M. Tombak and R. Siebert (1997): *Why Firms Form Research Joint Ventures: Theory and Evidence*; WZB Discussion Paper FS IV 97-6, WZB: Berlin.

Sandelin B. and N. Sarafoglou (2003): *Language and Scientific Publication Statistics: a Note*; Working Paper in Economics 109, University of Göteborg: Göteborg.

Sen A. and T. Smith (1995): *Gravity Models of Spatial Interaction Behaviour*; Advances in Spatial and Network Economics, Springer: Berlin, etc.

Sharp M. (1998): Competitiveness and Cohesion – are the two Compatible?; *Research Policy* 27(6), pp. 569-588.

Shen G. (1999): Estimating nodal attractions with exogenous spatial interaction and impedance data using the gravity model; *Papers in Regional Science* 78(2), pp. 213-220.

Teece D.J. (1982): Towards an Economic Theory of the Multiproduct Firm; *Journal of Economic Behaviour and Organization* 3, pp. 39-63.

Van der Meulen B. (2002): Europeanization of Research and the Role of Universities: an Organizational-Cultural Perspective; *Innovation* 15(4), pp. 341-355.

Vonortas N.S. (1989): *Inter-Firm Cooperation in Imperfectly Appropriable Research: Industry Performance and Welfare Implications*; Working Paper, Department of Economics, New York University: New York, NY.

Non-Academic Resources:

CIS (2004): *Official website of CIS (Community Innovation Survey)*;
http://www.cordis.lu/innovation-smes/src/cis.htm

COST (2004): *Official website of COST (European co-operation in the field of scientific and technical research)*;
http://cost.cordis.lu

CORDIS (2004): *Official website of CORDIS (Community Research & Development Information Service)*;
http://www.cordis.lu

Ecotec (2003): *Ex-Post Evaluation of Objective 1 1994-1999, A Final Report to the European Commission*; Ecotec Research & Consulting Ltd.: Birmingham.

Eureka (2003): *Annual report 2002 – 2003 Statistics*;
http://www.eureka.be/ifs/files/jsp-in/eureka/ifs/jsps/arStats.htm

European Commission (1998): *Research and Technological Development Activities of the European Union, 1998 Annual Report*; COM(98) 439, July 15th 1998.

European Commission (2001b): *Manual Of Proposal Evaluation Procedures*; Fifth Framework Programme;
http://www.cordis.lu/fp5/src/evalman.htm

European Commission (2001c): *Research and technological development activities of the European Union, 2001 Annual Report*; COM(2001) 756, December 12th 2001.

European Union (1992): *Treaty on European Union*; Office for Official Publications of the European Communities: Luxembourg.

European Union (1997): *A New Treaty for Europe*; Office for Official Publications of the European Communities: Luxembourg.

European Union (2003): *Verordnung (EG) Nr. 1059/2003 des europäischen Parlaments und des Rates vom 26. Mai 2003 über die Schaffung einer gemeinsamen Klassifikation der Gebietseinheiten für die Statistik (NUTS)*; May 26th 2003.

ESPRIT (1997): *Call on R&D Tasks, Two-Step Evaluation*, February 1997; http://www.cordis.lu/esprit/src/ind-2st.htm

JRC (2004): *Official website of the Joint Research Centre*;
http://www.jrc.cec.eu.int

Data Sources:

Eurostat (2003): Structural Indicators Database; Statistical Office of the European Communities (Eurostat): Luxembourg.

Amadeus (2003): Amadeus Business Intelligence; Amadeus Software: Leafield.

ARCS (2003), ARCS Data Base on European Framework Programme Collaboration ; Department of Technology Policy, Austrian Research Centers Seibersdorf: Seibersdorf.

European Commission (2001a): *Les Européens et les Langues, Eurobarometre 54 Special*; INRA, DG XXII de l'Education et de la Culture: Brussel-Bruxelles.

European Commission (2002): *Key Data on Education in Europe* 2002;

Software Used in Empirical Analysis:

ArcView (2002): *ArcView 3.2*; Environmental Systems Research Institute, Inc. (ESRI): Redlands, CA.

EEA (2004): GetMap Servlet; European Environmental Agency: Bruxelles.
http://map.eea.eu.int/getmap.asp

EViews (1997): *EViews 3.1*; Quantitative Micro Software (QMS): Irvine, CA.

Excel (1997): *Excel 97*; Microsoft Corporation: Redmond, WA.

Stata (2004): *Stata 7*; StataCorp LP: College Station, TX

APPENDIX A: COMPLEMENTING TABLES

Table A.1: NUTS-1 regions and identifiers

Code	Country (NUTS-0)	NUTS-1 Name	Largest City	Code	Country (NUTS-0)	NUTS-1 Name	Largest City
AT1	ÖSTERREICH	OSTÖSTERREICH	Wien	FR4	FRANCE	EST	Strasbourg
AT2	ÖSTERREICH	SÜDÖSTERREICH	Graz	FR5	FRANCE	OUEST	Nantes
AT3	ÖSTERREICH	WESTÖSTERREICH	Linz	FR6	FRANCE	SUD-OUEST	Toulouse
BE1	BELGIQUE-BELGIË	RÉGION DE BRUXELLES-CAPITALE BRUSSELS HOOFD-STEDELIJK GEWEST	Bruxelles /Brussel	FR7	FRANCE	CENTRE-EST	Lyon
BE2	BELGIQUE-BELGIË	VLAAMS GEWEST	Antwerpen	FR8	FRANCE	MÉDITERRANÉE	Marseille
BE3	BELGIQUE-BELGIË	RÉGION WALLONNE	Charleroi	GR1	ΕΛΛΑΔΑ	ΒΟΡΕΙΑ ΕΛΛΑΔΑ	Θεσσαλονίκη /Thessaloniki
DE1	DEUTSCHLAND	BADEN-WÜRTTEMBERG	Stuttgart	GR2	ΕΛΛΑΔΑ	ΚΕΝΤΡΙΚΗ ΕΛΛΑΔΑ	Πατραι/Patrai
DE2	DEUTSCHLAND	BAYERN	München	GR3	ΕΛΛΑΔΑ	ΑΤΤΙΚΗ	Αθήνα/Athina
DE3	DEUTSCHLAND	BERLIN	Berlin	GR4	ΕΛΛΑΔΑ	ΝΗΣΙΑ ΑΙΓΑΙΟΥ, ΚΡΗΤΗ	Ηράκλειο /Iraklion
DE4	DEUTSCHLAND	BRANDENBURG	Potsdam	IE0	IRELAND	IRELAND	Dublin
DE5	DEUTSCHLAND	BREMEN	Bremen	ITC	ITALIA	NORD-OVEST	Milano
DE6	DEUTSCHLAND	HAMBURG	Hamburg	ITD	ITALIA	NORD-EST	Bologna
DE7	DEUTSCHLAND	HESSEN	Frankfurt a.M.	ITE	ITALIA	CENTRO (I)	Roma
DE8	DEUTSCHLAND	MECKLENBURG-VORPOMMERN	Rostock	ITF	ITALIA	SUD	Napoli
DE9	DEUTSCHLAND	NIEDERSACHSEN	Hannover	ITG	ITALIA	ISOLE	Palermo
DEA	DEUTSCHLAND	NORDRHEIN-WESTFALEN	Köln	LU0	LUXEMBOURG (GRAND-DUCHÉ)	LUXEMBOURG (GRAND-DUCHÉ)	Luxembourg
DEB	DEUTSCHLAND	RHEINLAND-PFALZ	Mainz	NL1	NEDERLAND	NOORD-NEDERLAND	Groningen
DEC	DEUTSCHLAND	SAARLAND	Saarbrücken	NL2	NEDERLAND	OOST-NEDERLAND	Apeldoorn
DED	DEUTSCHLAND	SACHSEN	Leipzig	NL3	NEDERLAND	WEST-NEDERLAND	Amsterdam
DEE	DEUTSCHLAND	SACHSEN-ANHALT	Halle a.d. Saale	NL4	NEDERLAND	ZUID-NEDERLAND	Eindhoven
DEF	DEUTSCHLAND	SCHLESWIG-HOLSTEIN	Kiel	PT1	PORTUGAL	CONTINENTE	Lisboa
DEG	DEUTSCHLAND	THÜRINGEN	Erfurt	SE0	SVERIGE	SVERIGE	Stockholm
DK0	DANMARK	DANMARK	København	UKC	UNITED KINGDOM	NORTH EAST	Sunderland
ES1	ESPAÑA	NOROESTE	Vigo	UKD	UNITED KINGDOM	NORTH WEST	Liverpool
ES2	ESPAÑA	NORESTE	Zaragoza	UKE	UNITED KINGDOM	YORKSHIRE AND THE HUMBER	Leeds
ES3	ESPAÑA	COMUNIDAD DE MADRID	Madrid	UKF	UNITED KINGDOM	EAST MIDLANDS	Leicester
ES4	ESPAÑA	CENTRO (E)	Valladolid	UKG	UNITED KINGDOM	WEST MIDLANDS	Birmingham
ES5	ESPAÑA	ESTE	Barcelona	UKH	UNITED KINGDOM	EAST OF ENGLAND	Cambridge
ES6	ESPAÑA	SUR	Sevilla	UKI	UNITED KINGDOM	LONDON	London
ES7	ESPAÑA	CANARIAS	Las Palmas	UKJ	UNITED KINGDOM	SOUTH EAST	Brighton
FI1	SUOMI/FINLAND	MANNER-SUOMI	Helsinki /Helsingfors	UKK	UNITED KINGDOM	SOUTH WEST	Bristol
FR1	FRANCE	ÎLE-DE-FRANCE	Paris	UKL	UNITED KINGDOM	WALES	Cardiff
FR2	FRANCE	BASSIN PARISIEN	Le Havre	UKM	UNITED KINGDOM	SCOTLAND	Glasgow
FR3	FRANCE	NORD - PAS-DE-CALAIS	Lille	UKN	UNITED KINGDOM	NORTHERN IRELAND	Belfast

Source for NUTS-Codes and Labels: European Union (2003); “Largest City” is added in order to facilitate recognition, based on author’s assessment.

Table A.2: COMMUNITY RESEARCH COMMITMENTS OVER THE PERIOD 1984-2002 (CONSTANT PRICES 2000)

YEARS	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	TOTAL
FP1 1984-87	986.7	1153.8	1326.3	1030.5	369.9	136.4	6.3													5009.9
FP2 1987-91				276.2	1149.8	1675.2	2063.2	1561.1	274.2	17.3	4.5	0.2								7021.7
FP3 1990-94								363.6	2565.9	2435	2315.7	1.1								7681.3
FP4 1994-98*												3385.4	3465.4	3727.9	3679.6					14258.3
FP5 1998-02																3413.9	3633.7	3850.7	3906.6	14804.9
RTD PROGRAMMES	986.7	1153.8	1326.3	1306.7	1519.7	1811.6	2069.5	1924.7	2840.1	2452.3	2320.2	3386.7	3465.4	3727.9	3679.6	3413.9	3633.7	3850.7	3906.6	48776.1
APAS				72.5	80.3	94.2	146.1	207.4	366.3	515.5	657.2	2.4								2141.9
RTD+APAS	986.7	1153.8	1326.3	1379.2	1600	1905.8	2215.6	2132.1	3206.4	2967.8	2977.4	3389.1	3465.4	3727.9	3679.6	3413.9	3633.7	3850.7	3906.6	50918
SPRINT							20.7	19.7	20.2											60.6
ECSC							22.6	21.5	20.8	20.5	20.1									105.5
80% of THERMIE							46.5	145.5	153.1	163	167.4									675.5
Total Research**	**986.7**	**1153.8**	**1326.3**	**1379.2**	**1600**	**1905.8**	**2305.4**	**2318.8**	**3400.5**	**3151.3**	**3164.9**	**3389.1**	**3465.4**	**3727.9**	**3679.6**	**3413.9**	**3633.7**	**3850.7**	**3906.6**	**51759.6**

EC BUDGET	48095	46978	54388	56376	61106	57448	58213	68932	72722	79344	75780	85533	90247	90939	90981	94049	91667	92879	95907	1411584
RTD % of Budget	2.1	2.5	2.4	2.3	2.5	3.2	3.6	2.8	3.9	3.1	3.1	4.0	3.8	4.1	4.0	3.6	4.0	4.1	4.1	3.5
Research % of Budget	2.1	2.5	2.4	2.4	2.6	3.3	4.0	3.4	4.7	4.0	4.2	4.0	3.8	4.1	4.0	3.6	4.0	4.1	4.1	3.7
Deflation factors	0.601	0.637	0.659	0.681	0.705	0.741	0.774	0.814	0.842	0.854	0.870	0.881	0.910	0.935	0.951	0.974	1.000	1.018	1.038	-
Annual inflation (%)		6.0	3.5	3.3	3.6	5.1	4.5	5.2	3.5	1.4	1.9	1.3	3.3	2.7	1.7	2.4	2.7	1.8	2.0	-

*The amounts for FP 1994-98 are those adopted after the EU enlargement of 1995

**RTD + THERMIE + ECSC + SPRINT + APAS

Acronyms: FP - Framework Programme; RTD - Research & Technology Development; APAS - de preparation, d'accompagnement et de suivi [preparing, accompanying and follow-up measures]; SPRINT - Strategic Programme for Innovation and Technology Transfer; ECSC – European Coal and Steel Research; THERMIE - Technologies Européennes pour la Maîtrise de l'Energie [European Technologies for Mastering Energy]; EC – European Community;

Source: Adapted from European Commission (2001c), p. 53

Table A.3: Regional distribution of FP4 participants according to organisation type, prime contractors and participation in ICT.

	EDU	IND	ROR+ GOV	OTH	Total	%Pri	%IND	%ICT
AT1	300	200	188	51	739	18%	27%	12%
AT2	84	116	68	5	273	14%	42%	15%
AT3	81	163	34	12	290	13%	56%	19%
BE1	230	208	163	101	702	23%	30%	17%
BE2	554	514	290	49	1409	19%	36%	16%
BE3	339	217	71	12	639	19%	34%	10%
DE1	564	731	502	92	1890	21%	39%	16%
DE2	310	769	388	74	1541	17%	50%	24%
DE3	210	153	220	24	607	19%	25%	14%
DE4	28	80	78	8	194	16%	41%	8%
DE5	105	62	82	17	266	12%	23%	18%
DE6	125	120	119	13	377	16%	32%	11%
DE7	171	274	91	28	564	17%	49%	11%
DE8	32	32	13	2	79	18%	41%	4%
DE9	235	275	247	22	779	16%	35%	9%
DEA	515	718	455	99	1787	17%	40%	15%
DEB	67	104	85	11	267	18%	39%	11%
DEC	19	31	30	9	89	18%	35%	12%
DED	79	80	66	16	241	19%	33%	5%
DEE	47	20	31	3	101	15%	20%	12%
DEF	89	113	66	23	291	18%	39%	7%
DEG	23	56	44	16	139	23%	40%	27%
DK0	535	588	644	111	1878	17%	31%	9%
ES1	89	88	43	5	226	13%	39%	6%
ES2	61	295	214	30	600	13%	49%	17%
ES3	338	622	454	20	1435	16%	43%	22%
ES4	70	66	41	10	187	21%	35%	4%
ES5	491	500	375	30	1396	15%	36%	12%
ES6	154	94	157	4	409	20%	23%	3%
ES7	27	10	36	0	73	11%	14%	7%
FI1	451	444	574	94	1589	13%	28%	12%
FR1	406	1933	1360	115	3831	23%	50%	22%
FR2	97	192	170	16	475	19%	40%	7%
FR3	60	121	94	17	292	12%	41%	14%
FR4	149	140	196	21	506	21%	28%	5%
FR5	126	141	226	19	512	16%	28%	15%
FR6	172	317	252	23	764	22%	41%	10%
FR7	253	408	393	46	1100	21%	37%	10%

	EDU	IND	ROR+ GOV	OTH	Total	%Pri	%IND	%ICT
FR8	120	263	463	48	895	24%	29%	10%
GR1	190	112	60	22	384	14%	29%	6%
GR2	117	60	43	10	230	18%	26%	14%
GR3	461	675	325	88	1549	15%	44%	20%
GR4	66	25	141	4	236	14%	11%	11%
IE0	501	417	225	34	1240	14%	34%	17%
ITC	401	1086	935	76	2498	14%	43%	19%
ITD	375	380	315	51	1121	15%	34%	11%
ITE	479	519	496	57	1551	18%	33%	16%
ITF	196	103	197	6	502	15%	21%	6%
ITG	79	26	58	3	166	12%	16%	3%
LU0	3	62	39	14	119	20%	52%	19%
NL1	114	100	14	13	241	26%	41%	11%
NL2	348	279	404	63	1096	22%	25%	5%
NL3	740	593	1021	156	2513	22%	24%	8%
NL4	155	396	77	35	665	20%	60%	15%
PT1	526	516	553	43	1659	8%	31%	12%
SE0	1182	922	547	62	2754	13%	33%	10%
UKC	182	103	21	2	308	27%	33%	16%
UKD	329	282	89	24	724	20%	39%	11%
UKE	376	184	72	10	642	26%	29%	10%
UKF	227	287	110	7	631	18%	45%	7%
UKG	217	359	71	25	672	24%	53%	9%
UKH	480	365	405	34	1284	32%	28%	12%
UKI	857	429	268	81	1635	30%	26%	16%
UKJ	694	708	556	62	2020	26%	35%	15%
UKK	196	360	110	15	681	25%	53%	16%
UKL	171	103	24	8	306	23%	34%	6%
UKM	508	176	250	10	944	28%	19%	7%
UKN	92	40	14	3	149	17%	27%	9%
FI2	1	0	0	0	1	0%	0%	0%
FR9	0	4	6	0	10	30%	40%	0%
PT2	5	2	2	0	9	0%	22%	0%
PT3	2	2	3	0	7	14%	29%	0%
Other	1534	818	1834	269	4486	4%	18%	10%
Total	20537	22536	19466	2806	67682	17%	33%	13%

Figures are calculated from the ARCS (2003) database: **EDU** – Higher Education Institution. **IND** – Private Firm. **ROR** – Research Organisation (public, private or mixed). **GOV** – Governmental Body. **OTH** - other organisation types (consultancy or private non-commercial). %**Pri** - Percentage of prime contractors among FP4 participants in the region. **%IND** – Percentage of private firms among FP4 participants in the region. **%ICT** – Percentage of FP4 participants involved into information or communication technologies (ICT) key actions (ESPRIT + ACTS). **Total** differs from the sums of the columns or rows for the participants not attributable to a specific category.

Regions AT1 to UKN are the regions further analysed in this study. FI2 (Finnish Åland Islands), FR9 (French Overseas Départements), PT2 (Azores) and PT3 (Madeira) have been omitted for low collaboration numbers (compare section 5.1). "Other" are regions from non-EU-15 countries, mainly in Central and Eastern Europe.

Table A.4: NUTS-1 regions ranked by total implied distances

1) NUTS-1 region	2) Coord. 1 [1]	3) Coord. 2 [1]	4) Rowsum implied log-distances [2]	5) Inverse 4)	6) Rank 4)	7) Distance to origin[3]	8) No. Total Collaborations	9) Degree: 8) / Sum of 8) [4]	10) Rank 7)
FR1	-0.2273	-0.0725	72.6	0.0138	1	0.96	27,901	0.0632	1
SE0	-0.2731	-0.0755	74.9	0.0133	4	0.97	20,806	0.0471	2
ITC	-0.1324	-0.0649	75.0	0.0133	5	0.91	19,467	0.0441	3
NL3	-0.2399	-0.0273	74.6	0.0134	3	0.90	18,352	0.0416	4
UKJ	-0.1342	-0.0755	73.9	0.0135	2	0.81	15,127	0.0343	5
DK0	-0.2051	0.0119	78.5	0.0127	12	1.26	14,285	0.0324	6
DEA	-0.1295	-0.0194	77.6	0.0129	10	1.08	13,824	0.0313	7
DE1	-0.1793	-0.0297	76.3	0.0131	6	1.03	13,427	0.0304	8
FI1	-0.2508	-0.0658	78.8	0.0127	14	1.24	13,177	0.0299	9
PT1	-0.1954	-0.0703	77.7	0.0129	11	1.18	13,114	0.0297	10
DE2	-0.0071	-0.0559	77.3	0.0129	9	1.12	12,196	0.0276	11
ITE	-0.1464	-0.0340	78.5	0.0127	13	1.07	11,963	0.0271	12
GR3	-0.2195	0.0045	79.2	0.0126	15	1.17	11,334	0.0257	13
ES3	-0.1640	0.0518	80.0	0.0125	16	1.40	11,332	0.0257	14
UKI	-0.1590	-0.0753	77.0	0.0130	8	0.92	11,298	0.0256	15
BE2	-0.0625	0.0080	80.2	0.0125	17	1.06	10,708	0.0243	16
ES5	-0.1225	-0.1791	84.0	0.0119	24	1.55	10,051	0.0228	17
IE0	-0.1723	-0.0680	81.4	0.0123	20	1.30	9,140	0.0207	18
UKH	-0.1069	-0.1205	76.4	0.0131	7	1.03	8,889	0.0201	19
ITD	-0.2163	-0.0162	81.5	0.0123	21	1.39	8,547	0.0194	20
FR7	0.2199	-0.0634	84.2	0.0119	25	1.35	8,382	0.0190	21
NL2	-0.0977	-0.1011	80.5	0.0124	18	1.82	8,170	0.0185	22
FR8	-0.0368	-0.1211	86.0	0.0116	30	1.63	6,490	0.0147	23
UKM	-0.1574	-0.1892	85.2	0.0117	27	1.91	6,211	0.0141	24
DE9	-0.2039	-0.0245	81.0	0.0123	19	1.26	6,060	0.0137	25
FR6	0.0109	-0.0175	88.7	0.0113	36	1.64	6,029	0.0137	26
AT1	-0.0550	0.0715	88.1	0.0113	34	1.52	5,669	0.0128	27
BE1	1.0652	-0.0228	85.6	0.0117	28	1.65	5,619	0.0127	28
UKK	-0.0215	0.2609	85.8	0.0117	29	1.73	5,586	0.0127	29
UKF	1.3003	-0.1895	87.7	0.0114	33	1.70	5,526	0.0125	30
UKD	0.0939	-0.3280	84.4	0.0118	26	1.34	5,507	0.0125	31
UKG	0.1937	0.0888	90.3	0.0111	40	1.67	5,454	0.0124	32
UKE	-0.0806	-0.0196	86.4	0.0116	31	1.36	5,232	0.0119	33
NL4	1.4486	-0.3859	95.6	0.0105	44	2.25	5,078	0.0115	34
BE3	0.6117	0.2491	86.8	0.0115	32	1.74	4,936	0.0112	35
ES2	0.0543	-0.1053	97.5	0.0103	47	2.26	4,591	0.0104	36
DE3	0.4989	0.0143	83.9	0.0119	23	1.56	4,549	0.0103	37
DE7	-0.2799	-0.0344	83.7	0.0119	22	1.63	4,399	0.0100	38
ITF	0.3293	-0.1350	89.8	0.0111	39	1.83	4,186	0.0095	39
FR5	-0.2724	-0.0047	93.7	0.0107	41	2.20	4,180	0.0095	40
FR2	-0.1105	0.0888	88.4	0.0113	35	1.57	4,008	0.0091	41
FR4	1.1110	-0.0716	95.4	0.0105	43	1.95	3,516	0.0080	42
DE6	0.0911	0.0953	89.6	0.0112	38	2.03	3,380	0.0077	43
GR1	1.2614	1.2604	105.7	0.0095	56	2.40	3,035	0.0069	44
DEF	-0.3707	-0.4407	104.4	0.0096	53	2.60	2,842	0.0064	45

AT3	0.2377	1.4435	109.7	0.0091	57	2.87	2,669	0.0060	46
ES6	-0.1852	1.8222	99.1	0.0101	49	2.23	2,616	0.0059	47
FR3	0.8412	0.1539	100.8	0.0099	50	2.20	2,573	0.0058	48
UKC	-0.2483	0.3629	97.0	0.0103	46	1.94	2,520	0.0057	49
DE5	-0.3108	-0.0530	98.8	0.0101	48	2.11	2,467	0.0056	50
UKL	-0.3138	-0.1822	95.3	0.0105	42	2.07	2,262	0.0051	51
ES1	0.4285	-0.9431	113.2	0.0088	60	2.47	2,028	0.0046	52
DED	1.3182	-0.3415	104.3	0.0096	52	1.78	1,971	0.0045	53
GR2	0.4124	1.4614	104.8	0.0095	54	2.27	1,851	0.0042	54
AT2	0.3687	0.2405	105.0	0.0095	55	2.05	1,829	0.0041	55
DEB	0.5840	0.3523	104.0	0.0096	51	2.17	1,800	0.0041	56
GR4	0.0309	-0.5157	96.5	0.0104	45	1.82	1,787	0.0040	57
DE4	0.0014	0.0524	88.8	0.0113	37	1.52	1,725	0.0039	58
NL1	-0.6057	-0.0523	110.1	0.0091	59	2.47	1,568	0.0036	59
ES4	-0.0237	0.8863	115.9	0.0086	62	2.33	1,318	0.0030	60
ITG	0.6507	-0.2345	115.3	0.0087	61	2.78	1,272	0.0029	61
DEG	0.6499	-2.0362	133.9	0.0075	64	3.20	1,112	0.0025	62
UKN	0.2948	-0.3175	117.0	0.0085	63	2.27	1,021	0.0023	63
LU0	-0.8182	1.3062	144.6	0.0069	67	3.68	831	0.0019	64
DE8	-3.6050	0.2557	145.6	0.0069	68	3.52	699	0.0016	65
DEE	-0.7029	-3.1887	139.6	0.0072	65	3.05	656	0.0015	66
DEC	-0.4790	0.8793	110.0	0.0091	58	2.60	595	0.0013	67
ES7	-1.7859	-0.2530	144.4	0.0069	66	3.47	586	0.0013	68

1) Coordinates drawn from implied distances by procedure in (14). 2) Sum of 68 implied distances of the respective node to other nodes and itself. 3) Euklidian distance from origin to each 68-dimensional point defined by coordinate vectors. 4) Degree is defined as the sum of total cooperation links by the respective node, divided by the total sum of collaborative links (441,329).

Table A. 5: Implied distance matrix for EU-15 member states

	AT	**BE**	**DE**	**DK**	**ES**	**FI**	**FR**	**GR**	**IE**	**IT**	**LU**	**NL**	**PT**	**SE**	**UK**
AT	1	2.291	1.429	1.916	2.771	2.043	2.518	2.979	2.424	1.973	4.715	2.341	3.145	1.990	2.503
BE	2.291	1	1.393	1.642	1.979	1.836	1.212	2.036	1.732	1.574	2.259	1.248	1.710	1.666	1.460
DE	1.429	1.393	1	1.357	1.708	1.442	1.169	1.930	2.017	1.268	4.579	1.295	1.872	1.183	1.297
DK	1.916	1.642	1.357	1	2.125	1.220	1.679	1.900	1.563	1.756	3.869	1.218	1.853	0.883	1.261
ES	2.771	1.979	1.708	2.125	1	2.382	1.414	1.658	2.017	1.285	7.679	2.130	1.072	1.722	1.623
FI	2.043	1.836	1.442	1.220	2.382	1	2.002	1.811	1.604	1.881	4.691	1.625	2.063	0.986	1.659
FR	2.518	1.212	1.169	1.679	1.414	2.002	1	1.819	1.786	1.107	4.911	1.603	1.639	1.469	1.206
GR	2.979	2.036	1.930	1.900	1.658	1.811	1.819	1	2.066	1.275	5.504	2.147	1.567	2.156	1.681
IE	2.424	1.732	2.017	1.563	2.017	1.604	1.786	2.066	1	1.819	5.225	1.938	1.558	1.582	1.116
IT	1.973	1.574	1.268	1.756	1.285	1.881	1.107	1.275	1.819	1	9.341	1.767	1.539	1.555	1.320
LU	4.715	2.259	4.579	3.869	7.679	4.691	4.911	5.504	5.225	9.341	1	7.059	5.661	4.763	8.085
NL	2.341	1.248	1.295	1.218	2.130	1.625	1.603	2.147	1.938	1.767	7.059	1	2.159	1.318	1.292
PT	3.145	1.710	1.872	1.853	1.072	2.063	1.639	1.567	1.558	1.539	5.661	2.159	1	1.878	1.534
SE	1.990	1.666	1.183	0.883	1.722	0.986	1.469	2.156	1.582	1.555	4.763	1.318	1.878	1	1.183
UK	2.503	1.460	1.297	1.261	1.623	1.659	1.206	1.681	1.116	1.320	8.085	1.292	1.534	1.183	1

Table A.6: Indicators – Data Description

Identifier	Description	Sector	Unit			Plain				Log-transformed				Source
				Earliest obs	Latest obs	Mean	St.Dev.	J-B	Prob.	Mean	St.Dev.	J-B	Prob.	
EMPLOYEES	All NACE branches/ Employees /ESA95	-	1000 persons	1996	1998	1917.6	1357.4	30.3	0.000	7.2	0.77	2.9	0.239	Eurostat Yearbook
POP	Population at 1st January	-	1000 persons	1996	1998	5475.6	3650.9	18.1	0.000	8.3	0.75	5.8	0.055	Eurostat Yearbook
GDP_ME	GDP in Millions of ECU	-	Millions of ECU /ESA95	1996	1996	101526	84579	59.3	0.000	11.2	0.8	0.4	0.824	Eurostat Science
GDP_MP	GDP in Millions of PPP	-	Millions of PPP /ESA95	1996	1996	101525	80250	41.8	0.000	11.2	0.78	0.7	0.699	Eurostat Science
GDPP_INH	GDP in PPP per inhabitant	-	PPP per inhabitant /ESA95	1996	1996	18392.4	5518.9	91.2	0.000	9.7	0.26	6.0	0.050	Eurostat Science
GDPP_INHPC	GDP in PPP per inhabitant as percentage of EU-15 average	-	PPP/inhabitant in % of EU average /ESA95	1996	1996	99.4	29.8	91.3	0.000	4.5	0.26	6.2	0.045	Eurostat Science
GDPE_INH	GDP in ECU per inhabitant	-	ECU/inhabitant /ESA95	1996	1996	18522.5	6900.6	26.5	0.000	9.7	0.35	0.1	0.936	Eurostat Science
GDPE_INHPC	GDP in ECU per inhabitant as percentage of EU-15 average	-	ECU/inhabitant in % of EU average /ESA95	1996	1996	100.1	37.3	26.5	0.000	4.5	0.35	0.2	0.927	Eurostat Science
ENGLISH	Percentage of total number of students	-	ISCED1997: Upper sec. education/ level 3/ % students by modern language studied	1999	2000	0.71	0.18	2.2	0.330	0.53	0.1	1.2	0.547	European Comission (2001a)
TOT_HRST	TOT_INCL All NACE Rev. 1: codes 01 to 99ncluding not applicable and no answer	HRST Human Resources in Science and Technology	1000 Thousands	1996	1998	878.4	629.7	45.3	0.000	6.5	0.81	7.3	0.026	Eurostat Science
TOT_HRSTE	TOT_INCL All NACE Rev. 1: codes 01 to 99ncluding not applicable and no answer	HRST - Education	1000 Thousands	1996	1998	609	414.6	23.6	0.000	6.1	0.82	8.6	0.014	Eurostat Science
TOT_HRSTO	TOT_INCL All NACE Rev. 1: codes 01 to 99ncluding not applicable and no answer	HRST - Occupation	1000 Thousands	1996	1998	577.7	443.5	44.5	0.000	6	0.84	3.9	0.142	Eurostat Science
TOT_HRSTC	TOT_INCL All NACE Rev. 1: codes 01 to 99ncluding not applicable and no answer	HRST – Core	1000 Thousands	1996	1998	308.3	213.7	20.5	0.000	5.4	0.81	5.5	0.064	Eurostat Science
GQ_HRST	G_TO_Q Services	HRST Human Resources in Science and Technology	1000 Thousands	1996	1998	574.9	408	28.2	0.000	6	0.79	4.5	0.105	Eurostat Science
GQ_HRSTE	G_TO_Q Services	HRST - Education	1000 Thousands	1996	1998	362.5	243.2	17.4	0.000	5.6	0.8	6.9	0.031	Eurostat Science
GQ_HRSTO	G_TO_Q Services	HRST - Occupation	1000 Thousands	1996	1998	473.3	348.8	32.2	0.000	5.8	0.81	3.3	0.189	Eurostat Science
GQ_HRSTC	G_TO_Q Services	HRST – Core	1000 Thousands	1996	1998	260.9	174.5	16.7	0.000	5.3	0.78	5.7	0.057	Eurostat Science
KIS_HRST	KIS Knowledge Intensive Services (I61, 62,64 to J67, K70 to K74, M80, N85, 092)	HRST Human Resources in Science and Technology	1000 Thousands	1996	1998	389.8	274.3	12.4	0.002	5.6	0.81	4.4	0.113	Eurostat Science
KIS_HRSTE	KIS Knowledge Intensive Services (I61, 62,64 to J67, K70 to K74, M80, N85, 092)	HRST - Education	1000 Thousands	1996	1998	258.1	177.3	16.9	0.000	5.2	0.82	6.5	0.039	Eurostat Science
KIS_HRSTO	KIS Knowledge Intensive Services (I61, 62,64 to J67, K70 to K74, M80, N85, 092)	HRST - Occupation	1000 Thousands	1996	1998	342.3	241.7	12.7	0.002	5.5	0.81	3.8	0.147	Eurostat Science
KIS_HRSTC	KIS Knowledge Intensive Services (I61, 62,64 to J67, K70 to K74, M80, N85, 092)	HRST – Core	1000 Thousands	1996	1998	210.6	140	12.8	0.002	5	0.8	6.7	0.036	Eurostat Science
RDX_MEBIZ	R&D expenditure at the regional level	BES Business enterprise sector	Millions of ECU	1995	1998	1179.3	1715.6	232.0	0.000	6.1	1.5	6.4	0.040	Eurostat Science
RDX_MEGOV	R&D expenditure at the regional level	GOV Government sector	Millions of ECU	1995	1998	267.6	345.1	132.4	0.000	4.8	1.3	1.9	0.389	Eurostat Science
RDX_MEEDU	R&D expenditure at the regional level	HES Higher education sector	Millions of ECU	1995	1998	338.7	341.6	90.1	0.000	5.3	1.1	72.5	0.000	Eurostat Science

RDX_METOT3	R&D expenditure at the regional level	TOTAL excl Private Non-Profit	Millions of ECU	1995	1998	1785.7	2275	184.2	0.000	6.8	1.2	0.8	0.658	Eurostat Science
RDX_MPBIZ	R&D expenditure at the regional level	BES Business enterprise sector	Millions of PPS at 1995 prices	1995	1998	1109.9	1507.5	207.4	0.000	6.1	1.4	6.2	0.046	Eurostat Science
RDX_MPGOV	R&D expenditure at the regional level	GOV Government sector	Millions of PPS at 1995 prices	1995	1998	259.4	322.9	108.7	0.000	4.8	1.3	1.8	0.409	Eurostat Science
RDX_MPEDU	R&D expenditure at the regional level	HES Higher education sector	Millions of PPS at 1995 prices	1995	1998	328.6	297.2	39.1	0.000	5.3	1	100.8	0.000	Eurostat Science
RDX_MPTOT3	R&D expenditure at the regional level	TOTAL excl Private Non-Profit	Millions of PPS at 1995 prices	1995	1998	1698.1	1994.4	159.7	0.000	6.8	1.1	0.7	0.692	Eurostat Science
RDX_PCYBIZ	R&D expenditure at the regional level	BES Business enterprise sector	Percentage of GDP	1995	1998	0.95	0.71	13.5	0.001	0.6	0.34	1.5	0.481	Eurostat Science
RDX_PCYGOV	R&D expenditure at the regional level	GOV Government sector	Percentage of GDP	1995	1998	0.24	0.2	29.2	0.000	0.2	0.15	8.8	0.012	Eurostat Science
RDX_PCYEDU	R&D expenditure at the regional level	HES Higher education sector	Percentage of GDP	1995	1998	0.34	0.16	3.5	0.174	0.28	0.12	0.4	0.832	Eurostat Science
RDX_PCYTOT	R&D expenditure at the regional level	TOTAL excl Private Non-Profit	Percentage of GDP	1995	1998	1.5	0.87	8.6	0.013	0.88	0.33	1.3	0.534	Eurostat Science
RSTAFF_BIZ	R&D personnel at the regional level	BES Business enterprise sector	FTE Full time equivalent	1995	2000	12564.4	15714.5	161.6	0.000	8.7	1.3	11.3	0.004	Eurostat Science
RSTAFF_GOV	R&D personnel at the regional level	GOV Government sector	FTE Full time equivalent	1995	2000	3545.3	4142.5	71.0	0.000	7.5	1.2	1.4	0.505	Eurostat Science
RSTAFF_EDU	R&D personnel at the regional level	HES Higher education sector	FTE Full time equivalent	1995	2000	5262.5	4219.4	20.1	0.000	8.2	0.98	112.0	0.000	Eurostat Science
RSTAFF_TOT	R&D personnel at the regional level	All institutional sectors	FTE Full time equivalent	1995	2000	21653.3	22242.1	111.8	0.000	9.5	0.96	0.7	0.708	Eurostat Science
PAT_MLF	EPO Patent Applications	Total number of patent applications	Number of applications per million people in the labour force	1996	1997	223.7	223.9	125.5	0.000	4.8	1.2	9.0	0.011	Eurostat Science
PAT_M	EPO Patent Applications	Total number of patent applications	Total number of applications	1996	1997	691.1	1225.9	507.3	0.000	5.5	1.5	1.7	0.437	Eurostat Science
NBBIZ2	Amadeus 5,135,259 companies	<200	-	2003	2003	38995.4	41539.9	49.4	0.000	9.8	1.4	12.1	0.002	Amadeus
NBBIZ25	Amadeus 5,135,259 companies	200 to 500	-	2003	2003	468.3	456.7	57.4	0.000	5.7	1	6.4	0.040	Amadeus
NBBIZ510	Amadeus 5,135,259 companies	500 to 1000	-	2003	2003	151.6	167.3	104.5	0.000	4.5	1.1	2.7	0.257	Amadeus
NBBIZ10	Amadeus 5,135,259 companies	>1000	-	2003	2003	154.9	207.9	289.6	0.000	4.3	1.3	1.4	0.486	Amadeus
NBBIZ	Amadeus 5,135,259 companies	Total	-	2003	2003	77092.2	79964.5	276.3	0.000	10.7	1.1	18.5	0.000	Amadeus
TOPBIZ2	Amadeus Top 1.5 Million	<200	-	2003	2003	12296.4	12184.1	36.8	0.000	8.8	1.1	6.7	0.035	Amadeus
TOPBIZ25	Amadeus Top 1.5 Million	200 to 500	-	2003	2003	468.3	456.7	57.4	0.000	5.7	1	6.4	0.040	Amadeus
TOPBIZ510	Amadeus Top 1.5 Million	500 to 1000	-	2003	2003	151.6	167.3	104.5	0.000	4.5	1.1	2.7	0.257	Amadeus
TOPBIZ10	Amadeus Top 1.5 Million	>1000	-	2003	2003	154.9	207.9	289.6	0.000	4.3	1.3	1.4	0.486	Amadeus
TOPBIZ	Amadeus Top 1.5 Million	Total	-	2003	2003	16517.7	15409.7	29.3	0.000	9.2	1	1.8	0.400	Amadeus
PUBNAT_POP	Natural Science Articles / 1m inhabitants	National Average	-	2000	2000	2477.4	1002.5	10.5	0.005	7.7	0.39	0.31	0.855	Sandelin/ Sarafoglou (2003)
XPRES_BIZ	Million PPP per researcher in the sector	BES Business enterprise sector	RDX_MPBIZ /RSTAFF_BIZ	-	-	0.08	0.017	1.3	0.521	0.077	0.016	1.4	0.499	-
XPRES_GOV	Million PPP per researcher in the sector	GOV Government sector	RDX_MPGOV /RSTAFF_GOV	-	-	0.071	0.023	1.7	0.425	0.068	0.021	1.5	0.466	-
XPRES_EDU	Million PPP per researcher in the sector	HES Higher education sector	RDX_MPEDU /RSTAFF_EDU	-	-	0.06	0.017	0.9	0.623	0.058	0.016	0.9	0.649	-
XPRES_TOT	Million PPP per researcher in the sector	TOTAL excl Private Non-Profit	RDX_MPTOT3 /RSTAFF_TOT	-	-	0.07	0.017	2.5	0.286	0.068	0.016	2.7	0.255	-
RDXPC_BIZ	R&D Expenditure in % of total R&D Expenditure per Region	BES Business enterprise sector	RDX_MEBIZ /RDX_METOT3	-	-	0.55	0.2	3.9	0.139	0.43	0.13	6.1	0.046	-

RDXPC_GOV	R&D Expenditure in % of total R&D Expenditure per Region	GOV Government sector	RDX_MEGOV /RDX_METOT3	-	-	0.16	0.1	15.5	0.000	0.15	0.089	5.6	0.062	-
RDXPC_EDU	R&D Expenditure in % of total R&D Expenditure per Region	HES Higher education sector	RDX_MEEDU /RDX_METOT3	-	-	0.27	0.15	3.9	0.141	0.23	0.11	2.3	0.324	-
TOPFIRMSPC	Firms of Amadeus Top 1,500,000 >1000 employees in % of all Amadeus firms	-	TOPBIZ10 /NBBIZ	-	-	0.0022	0.0031	5801.8	0.000	0.0022	0.003	5718.6	0.000	-
EMP_PFIRM	Average Employees by Amadeus Firm	-	EMPLOYEES /NBBIZ	-	-	37.7	24	66.7	0.000	3.4	0.64	4.9	0.087	-
TOP10_ALLBIZ	Ratio big to small firms: Amadeus Top 1,500,000 firms >1000 employees / Amadeus firms <200 employees	-	TOPBIZ10 /NBBIZ2	-	-	0.01	0.018	1177.1	0.000	0.0097	0.017	1019.2	0.000	-
RDX_PTOP2_	Business R&D Expenditure divided by Top 1,500,000 firms over 200 employees	-	RDX_MEBIZ/ (TOPBIZ25 +TOPBIZ510 +TOPBIZ10)	-	-	1.3	0.98	6.5	0.040	0.78	0.4	2.9	0.232	-
TOP5_NB5BIZ	Top SMEs: Percentage of Top 1,500,000 companies among firms < 500 employees	-	(TOPBIZ2+TOPBIZ25) /(NBBIZ2+NBBIZ25)	-	-	0.42	0.2	8.6	0.013	0.34	0.13	7.3	0.026	-
RDXPNAT_BIZ	Regional Sector R&D Expenditure in % of National Sector R&D Expenditure	BES Business enterprise sector	RDX_MEBIZ(NUTS1) /RDX_MEBIZ(NUTS0)	-	-	0.22	0.28	46.8	0.000	0.17	0.2	28.1	0.000	-
RDXPNAT_GOV	Regional Sector R&D Expenditure in % of National Sector R&D Expenditure	GOV Government sector	RDX_MEGOV(NUTS1) /RDX_MEGOV(NUTS0)	-	-	0.22	0.29	43.5	0.000	0.17	0.2	28.6	0.000	-
RDXPNAT_EDU	Regional Sector R&D Expenditure in % of National Sector R&D Expenditure	HES Higher education sector	RDX_MEEDU(NUTS1) /RDX_MEEDU(NUTS0)	-	-	0.22	0.27	67.0	0.000	0.17	0.19	42.8	0.000	-
RDXPNAT_TOT	Regional Sector R&D Expenditure in % of National Sector R&D Expenditure	TOTAL excl Private Non-Profit	RDX_METOT3(NUTS1) /RDX_METOT3(NUTS0)	-	-	0.22	0.28	57.7	0.000	0.17	0.19	34.9	0.000	-
RSTAPNAT_BIZ	Regional Sector R&D Staff as Percentage of National Sector R&D Staff	BES Business enterprise sector	RSTAFF_BIZ(NUTS1) /RSTAFF_BIZ(NUTS0)	-	-	0.22	0.28	49.9	0.000	0.17	0.2	30.5	0.000	-
RSTAPNAT_GOV	Regional Sector R&D Staff as Percentage of National Sector R&D Staff	GOV Government sector	RSTAFF_GOV(NUTS1) /RSTAFF_GOV(NUTS0)	-	-	0.22	0.29	45.3	0.000	0.17	0.2	29.8	0.000	-
RSTAPNAT_EDU	Regional Sector R&D Staff as Percentage of National Sector R&D Staff	HES Higher education sector	RSTAFF_EDU(NUTS1) /RSTAFF_EDU(NUTS0)	-	-	0.22	0.27	69.9	0.000	0.17	0.19	44.6	0.000	-
RSTAPNAT_TOT	Regional Sector R&D Staff as Percentage of National Sector R&D Staff	All institutional sectors	RSTAFF_TOT(NUTS1) /RSTAFF_TOT(NUTS0)	-	-	0.22	0.27	63.2	0.000	0.17	0.19	38.8	0.000	-
PATPRX_BIZ	Million Patents per Biz expenditure	-	PAT_M /RDX_MPBIZ	-	-	0.63	0.46	84.4	0.000	0.45	0.24	13.7	0.001	-
PATPST_BIZ	Million Patents per Business researchers	-	PAT_M /RSTAFF_BIZ	-	-	0.048	0.032	146.0	0.000	0.046	0.029	109.2	0.000	-
GDPP_EMP	GDP per employee	-	GDP_MP /EMPLOYEES	-	-	18.4	5.5	92.4	0.000	2.9	0.25	7.1	0.029	-
DRXMP_BIZGOV	Biz over Govt R&D Expenditure	-	RDX_MPBIZ /RDX_MPGOV	-	-	9.3	23.1	5955.6	0.000	1.7	0.94	20.3	0.000	-
RSTAFF_PCLBIZ	Business researchers per employee	BES Business enterprise sector	RSTAFF_BIZ /EMPLOYEES	-	-	0.0057	0.0038	6.6	0.037	0.0056	0.0038	6.5	0.040	-
RSTAFF_PCLGOV	Government researchers per employee	GOV Government sector	RSTAFF_GOV /EMPLOYEES	-	-	0.0017	0.0015	44.6	0.000	0.0017	0.0014	44.3	0.000	-
RSTAFF_PCLEDU	HES researchers per employee	HES Higher education sector	RSTAFF_EDU /EMPLOYEES	-	-	0.0028	0.0012	2.0	0.367	0.0028	0.0012	2.0	0.374	-
RSTAFF_PCLTOT	Total researchers per employee	All institutional sectors	RSTAFF_TOT /EMPLOYEES	-	-	0.01	0.0046	6.1	0.047	0.01	0.0046	6.0	0.050	-
PAT_TOPBIZ	Patents per Top firm	-	PAT_M /TOPBIZ	-	-	0.041	0.039	66.1	0.000	0.039	0.037	53.1	0.000	-
PUBNAT_RST	Natural sciences publications per researcher (NUTS0)	National Average	PUBNAT_POP *POP/1000 /RSTAFF_TOT	-	-	0.66	0.25	0.2	0.901	0.49	0.15	2.9	0.232	-
GEODIST	Geographic great circle distance	-	kilometres	-	-	1405.7	890.2	291.8	0.000	6.936	1.117	19599	0.000	ARCVIEW

LANGSPOK	Dissimilarity in spoken language vectors	-	Cosine of national percentages of speakers of German, English, French	-	-	1.057	0.457	803.9	0.000	0.687	0.284	1056	0.000	European Commission (2001a)
LANGSTUD	Dissimilarity in vectors of language studied at ISCED-1997 level 3	-	Cosine of national percentages of speakers of EU-15 languages, Arabic, Chinese and Russion	-	-	0.876	0.391	236.2	0.000	0.602	0.247	335.2	0.000	European Commission (2002)
CULTDIM	Dissimilarity in Hofstede (1980) cultural dimensions	-	Cosine of the four dimensions	-	-	0.326	0.207	35.3	0.000	0.270	0.159	60.9	0.000	Hofstede
INDSTRUCT	Dissimilarity in Amadeus 5,135,259 companies size categories	-	Cosine of TOPBIZ2, TOPBIZ25, TOPBIZ510 and TOPBIZ10	-	-	0.052	0.056	1011.5	0.000	0.050	0.051	705.8	0.000	-
RDXDIFF	Dissimilarity in research expenditure per sector	-	Cosine of RDX_MPBIZ, RDX_MPEDU and RDX_MPGOV	-	-	0.267	0.170	177.6	0.000	0.228	0.130	81.3	0.000	-
RSTAFFDIFF	Dissimilarity in research staff per sector	-	Cosine of RSTAFF_BIZ, RSTAFF_EDU and RSTAFF_GOV	-	-	0.281	0.179	162.3	0.000	0.238	0.136	85.3	0.000	-
PATSTRUCT	Dissimilarity in EPO patent applications	-	Cosine of sectors A to H in EPO industry classification	-	-	0.553	0.211	1.4	0.506	0.431	0.138	30.9	0.000	Eurostat Science

Table A.7: Correlations of Log-transformed Mass Indicators with Collaboration Matrix Derivates

Identifier	Eigenvector 1st	2nd	3rd	4th	5th	6th	7th	8th	9th	Implied Mass	Row Sum
EMPLOYEES	0.66	0.13	-0.15	0.08	-0.10	0.05	0.16	0.05	0.11	0.66	0.67
POP	0.60	0.21	-0.06	0.06	-0.11	0.03	0.11	0.07	0.05	0.61	0.60
GDP_ME	0.71	0.11	-0.28	0.06	-0.10	0.11	0.08	0.14	0.09	0.69	0.71
GDP_MP	0.72	0.18	-0.18	0.07	-0.08	0.07	0.07	0.09	0.09	0.71	0.73
GDPP_INH	0.43	-0.06	-0.36	0.03	0.09	0.13	-0.09	0.08	0.11	0.35	0.42
GDPP_INHPC	0.43	-0.06	-0.36	0.03	0.09	0.13	-0.09	0.08	0.11	0.35	0.42
GDPE_INH	0.34	-0.19	-0.49	0.01	0.01	0.19	-0.05	0.19	0.10	0.26	0.33
GDPE_INHPC	0.34	-0.19	-0.49	0.01	0.01	0.19	-0.05	0.18	0.10	0.26	0.33
ENGLISH	-0.12	0.14	0.17	0.11	0.19	-0.16	0.22	-0.28	-0.15	-0.14	-0.12
TOT_HRST	0.68	0.05	-0.21	0.06	-0.06	0.06	0.09	0.06	0.11	0.67	0.68
TOT_HRSTE	0.66	0.03	-0.16	0.04	0.00	0.00	0.11	0.00	0.16	0.65	0.66
TOT_HRSTO	0.69	0.04	-0.24	0.07	-0.08	0.08	0.08	0.10	0.07	0.67	0.69
TOT_HRSTC	0.70	0.00	-0.17	0.05	0.01	0.01	0.10	0.03	0.13	0.69	0.70
GQ_HRST	0.71	0.04	-0.21	0.06	-0.03	0.06	0.07	0.07	0.08	0.69	0.71
GQ_HRSTE	0.70	0.01	-0.13	0.04	0.05	-0.02	0.09	0.01	0.13	0.69	0.70
GQ_HRSTO	0.70	0.04	-0.22	0.07	-0.06	0.09	0.06	0.10	0.07	0.68	0.70
GQ_HRSTC	0.71	0.01	-0.14	0.05	0.03	0.01	0.08	0.04	0.13	0.70	0.71
KIS_HRST	0.73	0.02	-0.19	0.07	0.01	0.03	0.07	0.06	0.06	0.71	0.73
KIS_HRSTE	0.72	0.01	-0.12	0.05	0.07	-0.04	0.10	0.01	0.12	0.70	0.72
KIS_HRSTO	0.73	0.03	-0.19	0.07	-0.02	0.06	0.06	0.08	0.05	0.70	0.73
KIS_HRSTC	0.72	0.01	-0.12	0.05	0.04	-0.02	0.09	0.03	0.11	0.71	0.72
RDX_MEBIZ	0.64	-0.07	-0.47	0.05	-0.04	0.06	0.15	0.08	0.05	0.58	0.63
RDX_MEGOV	0.63	0.04	-0.23	0.12	0.16	0.23	0.06	0.22	0.08	0.56	0.63
RDX_MEEDU	0.62	-0.13	-0.28	0.01	-0.10	0.11	-0.07	0.08	0.05	0.58	0.62
RDX_METOT3	0.70	-0.06	-0.46	0.06	-0.02	0.10	0.10	0.11	0.07	0.63	0.69
RDX_MPBIZ	0.67	-0.03	-0.44	0.06	-0.03	0.04	0.15	0.05	0.04	0.61	0.66
RDX_MPGOV	0.66	0.09	-0.17	0.13	0.18	0.22	0.05	0.19	0.08	0.59	0.65
RDX_MPEDU	0.64	-0.08	-0.20	0.02	-0.08	0.09	-0.08	0.05	0.04	0.60	0.64
RDX_MPTOT3	0.74	-0.01	-0.41	0.07	0.00	0.08	0.09	0.08	0.06	0.68	0.74
RDX_PCYBIZ	0.48	-0.19	-0.55	-0.01	0.04	0.03	0.20	-0.03	0.04	0.39	0.47
RDX_PCYGOV	0.30	-0.02	-0.15	0.13	0.35	0.16	-0.02	0.21	0.04	0.21	0.30
RDX_PCYEDU	0.17	-0.47	-0.16	0.00	0.00	0.10	-0.23	-0.02	0.00	0.10	0.16
RDX_PCYTOT	0.49	-0.25	-0.51	0.02	0.12	0.07	0.11	0.01	0.04	0.38	0.48
RSTAFF_BIZ	0.65	-0.05	-0.42	0.04	-0.06	0.04	0.16	0.04	0.03	0.60	0.64
RSTAFF_GOV	0.66	0.10	-0.07	0.12	0.14	0.25	-0.02	0.17	0.11	0.62	0.66
RSTAFF_EDU	0.66	-0.01	-0.06	0.01	-0.08	0.12	-0.12	0.00	0.07	0.64	0.66
RSTAFF_TOT	0.79	0.01	-0.32	0.05	-0.03	0.11	0.05	0.03	0.08	0.74	0.78
PAT_MLF	0.30	-0.23	-0.64	0.05	-0.09	0.08	0.07	0.15	-0.08	0.20	0.28
PAT_M	0.54	-0.11	-0.56	0.08	-0.12	0.09	0.11	0.15	-0.04	0.47	0.53
NBBIZ2	0.47	0.05	-0.32	0.01	-0.22	0.19	-0.05	0.31	0.27	0.47	0.47
NBBIZ25	0.71	0.03	-0.26	0.08	-0.05	0.08	0.10	0.03	0.11	0.68	0.70
NBBIZ510	0.73	-0.05	-0.28	0.07	-0.02	0.06	0.11	-0.04	0.16	0.70	0.72
NBBIZ10	0.74	-0.05	-0.26	0.08	0.02	0.04	0.08	-0.08	0.18	0.69	0.74
NBBIZ	0.63	-0.01	-0.19	0.07	0.02	-0.13	0.17	0.07	0.26	0.60	0.62
TOPBIZ2	0.59	0.12	-0.23	0.04	-0.20	0.20	-0.07	0.24	0.22	0.60	0.59
TOPBIZ25	0.71	0.03	-0.26	0.08	-0.05	0.08	0.10	0.03	0.11	0.68	0.70
TOPBIZ510	0.73	-0.05	-0.28	0.07	-0.02	0.06	0.11	-0.04	0.16	0.70	0.72
TOPBIZ10	0.74	-0.05	-0.26	0.08	0.02	0.04	0.08	-0.08	0.18	0.69	0.74
TOPBIZ	0.71	0.11	-0.10	0.05	-0.15	0.07	-0.02	0.12	0.30	0.73	0.71
PUBNAT_POP	0.23	-0.60	-0.14	0.01	0.13	-0.19	0.03	-0.06	-0.16	0.20	0.23
XPRES_BIZ	0.41	0.10	-0.28	0.11	0.16	0.01	0.05	0.09	0.11	0.35	0.41
XPRES_GOV	0.05	-0.02	-0.32	0.05	0.21	-0.09	0.27	0.10	-0.13	-0.05	0.04
XPRES_EDU	0.15	-0.24	-0.48	0.03	-0.06	-0.02	0.06	0.16	-0.11	0.07	0.14
XPRES_TOT	0.33	-0.07	-0.55	0.08	0.10	-0.01	0.18	0.20	-0.02	0.24	0.32
RDXPC_BIZ	0.29	-0.04	-0.43	0.01	-0.09	-0.08	0.28	-0.04	-0.01	0.27	0.29
RDXPC_GOV	-0.03	0.13	0.31	0.11	0.27	0.15	-0.16	0.18	0.02	-0.05	-0.03
RDXPC_EDU	-0.34	-0.06	0.38	-0.09	-0.06	0.01	-0.28	-0.05	0.00	-0.30	-0.34
TOPFIRMSPC	0.04	0.00	-0.09	0.02	0.00	0.15	-0.10	-0.13	0.00	0.02	0.03
EMP_PFIRM	0.33	-0.17	-0.15	0.03	0.16	-0.30	0.10	0.06	0.33	0.30	0.33
TOP10_ALLBIZ	0.02	-0.04	0.16	0.05	0.20	-0.21	0.17	-0.34	-0.07	0.02	0.02
RDX_PTOP2_	0.23	-0.10	-0.53	-0.03	-0.04	0.04	0.15	0.15	-0.05	0.17	0.22
TOP5_NB5BIZ	0.01	0.13	0.39	0.06	0.20	-0.09	0.03	-0.34	-0.32	0.04	0.02
RDXPNAT_BIZ	0.62	-0.28	0.22	-0.14	-0.11	0.14	-0.06	-0.22	0.07	0.70	0.63
RDXPNAT_GOV	0.60	-0.33	0.30	-0.08	0.07	0.16	-0.14	-0.15	0.07	0.67	0.61
RDXPNAT_EDU	0.56	-0.37	0.30	-0.18	-0.05	0.14	-0.12	-0.15	0.05	0.65	0.57
RDXPNAT_TOT	0.61	-0.32	0.26	-0.14	-0.06	0.14	-0.09	-0.20	0.07	0.70	0.62
RSTAPNAT_BIZ	0.61	-0.29	0.24	-0.15	-0.13	0.12	-0.06	-0.22	0.06	0.70	0.62
RSTAPNAT_GOV	0.59	-0.33	0.30	-0.07	0.08	0.17	-0.14	-0.15	0.06	0.66	0.60
RSTAPNAT_EDU	0.54	-0.37	0.32	-0.19	-0.05	0.13	-0.12	-0.16	0.04	0.64	0.55
RSTAPNAT_TOT	0.60	-0.33	0.28	-0.15	-0.06	0.13	-0.09	-0.19	0.05	0.68	0.61
PATPRX_BIZ	-0.23	-0.15	-0.28	0.04	-0.22	0.11	-0.06	0.18	-0.11	-0.28	-0.23
PATPST_BIZ	-0.10	-0.10	-0.36	0.08	-0.23	0.12	-0.06	0.19	-0.08	-0.17	-0.11
GDPP_EMP	0.43	-0.06	-0.36	0.04	0.09	0.13	-0.10	0.08	0.11	0.34	0.42
DRXMP_BIZGOV	0.07	-0.13	-0.35	-0.08	-0.25	-0.21	0.10	-0.16	-0.04	0.08	0.07
RSTAFF_PCLBIZ	0.48	-0.14	-0.58	-0.12	0.00	0.05	0.08	-0.04	0.05	0.39	0.47
RSTAFF_PCLGOV	0.33	0.03	0.00	0.08	0.29	0.25	-0.29	0.14	0.06	0.27	0.33
RSTAFF_PCLEDU	0.21	-0.26	0.11	-0.04	0.06	0.16	-0.53	-0.06	0.00	0.19	0.20
RSTAFF_PCLTOT	0.57	-0.18	-0.44	-0.09	0.13	0.15	-0.16	-0.02	0.06	0.47	0.56
PAT_TOPBIZ	0.10	-0.06	-0.56	0.13	-0.19	0.11	0.14	-0.06	-0.13	0.00	0.09

PUBNAT_RST	0.12	-0.06	0.39	-0.02	0.13	-0.19	-0.04	-0.21	-0.16	0.19	0.13

Table A.8: Monte Carlo stability test results for LL_IMPDIST regression on variable doubles (50 runs each)

Variable 1	Variable 2	mean f-stat	% insign p-val	Rank	CONST	LL_IMPDISTSF	Var 1	Var 2
INDSTRUCT	PATSTRUCT	0.948	0.84	1	-0.88	1.00	0.50	0.98
INDSTRUCT	RDXDIFF	0.977	0.82	2	0.00	1.00	0.56	1.00
RSTAFFDIFF	GEODIST	0.955	0.82	2	-0.16	1.00	1.00	1.00
RDXDIFF	GEODIST	0.963	0.82	2	-0.14	1.00	1.00	1.00
L_GEODIST	PATSTRUCT	1.060	0.80	5	-1.00	1.00	1.00	0.26
LANGSTUD	L_DIST2CORE	1.018	0.80	5	-1.00	1.00	1.00	0.36
L_GEODIST	CULTDIM	1.003	0.78	7	-1.00	1.00	1.00	0.02
LANGSPOK	PATSTRUCT	1.017	0.78	7	-1.00	1.00	1.00	0.96
LANGSTUD	PATSTRUCT	1.013	0.78	7	-1.00	1.00	1.00	0.72
L_GEODIST	RSTAFFDIFF	1.062	0.76	10	-1.00	1.00	1.00	0.98
LANGSPOK	INDSTRUCT	1.056	0.76	10	-1.00	1.00	1.00	0.24
LANGSPOK	CULTDIM	1.039	0.76	10	-1.00	1.00	1.00	0.80
RSTAFFDIFF	ENGLISH	1.053	0.76	10	0.04	1.00	1.00	-0.08
PATSTRUCT	L_DIST2CORE	1.040	0.76	10	-0.88	1.00	1.00	-0.02
RDXDIFF	ENGLISH	1.050	0.76	10	0.02	1.00	1.00	-0.04
CULTDIM	GEODIST	1.037	0.76	10	-0.84	1.00	0.92	1.00
L_GEODIST	RDXDIFF	1.059	0.74	17	-1.00	1.00	1.00	0.98
LANGSPOK	ENGLISH	1.058	0.74	17	-1.00	1.00	1.00	0.06
LANGSPOK	L_DIST2CORE	1.067	0.74	17	-1.00	1.00	1.00	0.00
RSTAFFDIFF	RDXDIFF	1.020	0.74	17	-0.02	1.00	0.56	0.00
PATSTRUCT	RDXDIFF	1.060	0.74	17	-0.86	1.00	0.98	1.00
RDXDIFF	CULTDIM	1.071	0.74	17	-1.00	1.00	1.00	1.00
ENGLISH	L_DIST2CORE	1.004	0.74	17	0.06	1.00	-0.10	-0.02
ENGLISH	GEODIST	1.038	0.74	17	0.00	1.00	-0.56	1.00
L_GEODIST	LANGSPOK	1.008	0.72	25	-1.00	1.00	1.00	0.84
L_GEODIST	INDSTRUCT	1.052	0.72	25	-1.00	1.00	1.00	0.02
LANGSTUD	CULTDIM	1.066	0.72	25	-1.00	1.00	1.00	0.60
RSTAFFDIFF	CULTDIM	1.047	0.72	25	-1.00	1.00	1.00	1.00
PATSTRUCT	GEODIST	1.007	0.72	25	-0.44	1.00	0.46	1.00
CULTDIM	ENGLISH	1.074	0.72	25	-1.00	1.00	1.00	-0.20
LANGSPOK	RDXDIFF	1.008	0.70	31	-0.98	1.00	1.00	1.00
LANGSTUD	RSTAFFDIFF	1.085	0.70	31	-1.00	1.00	1.00	1.00
LANGSTUD	RDXDIFF	1.067	0.70	31	-1.00	1.00	1.00	1.00
INDSTRUCT	CULTDIM	1.026	0.70	31	-1.00	1.00	0.14	1.00
RSTAFFDIFF	PATSTRUCT	1.023	0.70	31	-0.40	1.00	1.00	0.54
RSTAFFDIFF	L_DIST2CORE	1.041	0.70	31	0.00	1.00	1.00	-0.04
PATSTRUCT	ENGLISH	1.039	0.70	31	-0.64	1.00	0.98	-0.10
L_GEODIST	L_DIST2CORE	1.089	0.68	38	-1.00	1.00	1.00	0.00
LANGSPOK	RSTAFFDIFF	1.053	0.68	38	-1.00	1.00	1.00	1.00
LANGSTUD	INDSTRUCT	1.140	0.68	38	-1.00	1.00	1.00	0.40
INDSTRUCT	RSTAFFDIFF	1.105	0.68	38	-0.02	1.00	0.44	1.00
INDSTRUCT	ENGLISH	1.032	0.68	38	0.12	1.00	0.74	-0.12
INDSTRUCT	GEODIST	1.095	0.68	38	-0.06	1.00	0.20	1.00
PATSTRUCT	CULTDIM	1.133	0.68	38	-1.00	1.00	0.98	1.00
L_GEODIST	LANGSTUD	1.097	0.66	45	-1.00	1.00	1.00	0.58
CULTDIM	L_DIST2CORE	1.137	0.66	45	-1.00	1.00	1.00	0.02

L_DIST2CORE	GEODIST	1.055	0.66	45	-0.10	1.00	0.00	1.00
LANGSPOK	LANGSTUD	1.146	0.64	48	-1.00	1.00	0.80	1.00
LANGSPOK	GEODIST	1.181	0.64	48	-1.00	1.00	1.00	1.00
LANGSTUD	ENGLISH	1.138	0.64	48	-1.00	1.00	1.00	0.10
INDSTRUCT	L_DIST2CORE	1.120	0.64	48	0.02	1.00	0.52	0.00
L_GEODIST	GEODIST	1.128	0.62	52	-1.00	1.00	1.00	-1.00
RDXDIFF	L_DIST2CORE	1.114	0.60	53	0.00	1.00	1.00	-0.02
L_GEODIST	ENGLISH	1.118	0.58	54	-1.00	1.00	1.00	-0.72
LANGSTUD	GEODIST	1.119	0.58	54	-1.00	1.00	1.00	1.00

Each variable double (Var1, Var2) was included into a regression on *LL_IMPDIST* along with a constant and *LL_IMPDISTSF* (implied distance row sum model fitted values). Each pair was evaluated in a “Chow forecast test” between two randomly drawn data sub-samples with m(m+1)/2=595 observations. “Mean f-stats” depicts each double’s average Chow forecast statistic. “% insigni. pval.” is the number of cases (as share of total) where the Chow f-statistic implied a type I error probability of more than 0.05. Equivalently, “% signi. t-Stat” exhibits the share of cases in which each variable coefficient proved significant (t-stat prob value lower than 0.05). “Rank” ranks pairs according to stability (largest “% insigni. p-val”).

Table A.9: Estimating Implied distance with regional fixed effect dummies

Dependent Variable: LL_IMPDIST
Method: Least Squares
Included observations: 2346

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.205769	0.032298	6.370907	0.0000
DUMMY_AT1	0.333255	0.022923	14.53814	0.0000
DUMMY_AT2	0.387939	0.022923	16.92373	0.0000
DUMMY_AT3	0.454157	0.022923	19.81250	0.0000
DUMMY_BE1	0.217745	0.022923	9.499065	0.0000
DUMMY_BE2	0.238066	0.022923	10.38557	0.0000
DUMMY_BE3	0.247573	0.022923	10.80030	0.0000
DUMMY_DE1	0.179009	0.022923	7.809241	0.0000
DUMMY_DE2	0.180065	0.022923	7.855292	0.0000
DUMMY_DE3	0.203294	0.022923	8.868628	0.0000
DUMMY_DE4	0.250238	0.022923	10.91657	0.0000
DUMMY_DE5	0.387636	0.022923	16.91054	0.0000
DUMMY_DE6	0.244012	0.022923	10.64497	0.0000
DUMMY_DE7	0.222255	0.022923	9.695825	0.0000
DUMMY_DE8	0.577098	0.022923	25.17573	0.0000
DUMMY_DE9	0.221200	0.022923	9.649810	0.0000
DUMMY_DEA	0.217237	0.022923	9.476908	0.0000
DUMMY_DEB	0.359388	0.022923	15.67821	0.0000
DUMMY_DEC	0.419148	0.022923	18.28523	0.0000
DUMMY_DED	0.402135	0.022923	17.54304	0.0000
DUMMY_DEE	0.586006	0.022923	25.56435	0.0000
DUMMY_DEF	0.459896	0.022923	20.06283	0.0000
DUMMY_DEG	0.585361	0.022923	25.53621	0.0000
DUMMY_DK0	0.231901	0.022923	10.11661	0.0000
DUMMY_ES1	0.481668	0.022923	21.01263	0.0000
DUMMY_ES2	0.458615	0.022923	20.00696	0.0000
DUMMY_ES3	0.250081	0.022923	10.90970	0.0000
DUMMY_ES4	0.523485	0.022923	22.83691	0.0000
DUMMY_ES5	0.313611	0.022923	13.68120	0.0000
DUMMY_ES6	0.344294	0.022923	15.01974	0.0000
DUMMY_ES7	0.616636	0.022923	26.90060	0.0000
DUMMY_FI1	0.260559	0.022923	11.36684	0.0000
DUMMY_FR1	0.159721	0.022923	6.967808	0.0000
DUMMY_FR2	0.309380	0.022923	13.49660	0.0000
DUMMY_FR3	0.367998	0.022923	16.05382	0.0000
DUMMY_FR4	0.342551	0.022923	14.94369	0.0000
DUMMY_FR5	0.330410	0.022923	14.41406	0.0000
DUMMY_FR6	0.298461	0.022923	13.02029	0.0000
DUMMY_FR7	0.270643	0.022923	11.80675	0.0000
DUMMY_FR8	0.299188	0.022923	13.05198	0.0000
DUMMY_GR1	0.425118	0.022923	18.54568	0.0000
DUMMY_GR2	0.348741	0.022923	15.21372	0.0000
DUMMY_GR3	0.243806	0.022923	10.63599	0.0000
DUMMY_GR4	0.330498	0.022923	14.41788	0.0000
DUMMY_IE0	0.263014	0.022923	11.47391	0.0000
DUMMY_ITC	0.174118	0.022923	7.595842	0.0000
DUMMY_ITD	0.253939	0.022923	11.07803	0.0000
DUMMY_ITE	0.226972	0.022923	9.901576	0.0000
DUMMY_ITF	0.314956	0.022923	13.73986	0.0000
DUMMY_ITG	0.467130	0.022923	20.37845	0.0000
DUMMY_LU0	0.647118	0.022923	28.23034	0.0000
DUMMY_NL1	0.443731	0.022923	19.35763	0.0000
DUMMY_NL2	0.222523	0.022923	9.707492	0.0000
DUMMY_NL3	0.166658	0.022923	7.270405	0.0000
DUMMY_NL4	0.350413	0.022923	15.28666	0.0000
DUMMY_PT1	0.240093	0.022923	10.47399	0.0000
DUMMY_SE0	0.206082	0.022923	8.990253	0.0000
DUMMY_UKC	0.361946	0.022923	15.78982	0.0000
DUMMY_UKD	0.241016	0.022923	10.51427	0.0000
DUMMY_UKE	0.288898	0.022923	12.60311	0.0000
DUMMY_UKF	0.254398	0.022923	11.09805	0.0000
DUMMY_UKG	0.324225	0.022923	14.14423	0.0000
DUMMY_UKH	0.144653	0.022923	6.310433	0.0000
DUMMY_UKI	0.175347	0.022923	7.649474	0.0000
DUMMY_UKJ	0.127491	0.022923	5.561743	0.0000
DUMMY_UKK	0.242575	0.022923	10.58227	0.0000
DUMMY_UKL	0.326341	0.022923	14.23653	0.0000
DUMMY_UKM	0.290250	0.022923	12.66210	0.0000
DUMMY_UKN	0.481770	0.022923	21.01711	0.0000

R-squared	0.647961	Mean dependent var	0.838110
Adjusted R-squared	0.637447	S.D. dependent var	0.220345
S.E. of regression	0.132675	Akaike info criterion	-1.172856
Sum squared resid	40.08133	F-statistic	61.63275
Log likelihood	1444.760	Prob(F-statistic)	0.000000

Table A.10: Estimation results for implied distance modelling with estimated row sums, bilateral factors and regional dummies

Dependent Variable: LL_IMPDIST
Method: Least Squares
Included observations: 2278

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.122380	0.311291	-6.818000	0.0000
L_IMPDISTSF	0.238103	0.027306	8.719698	0.0000
RDXDIFF	0.153921	0.019288	7.980209	0.0000
SQRT_GEODIST	0.004904	0.000354	13.86077	0.0000
INTRAROMANIC	-0.083683	0.010866	-7.701067	0.0000
…				
…				
R-squared	0.717575		Mean dependent var	0.838110
Adjusted R-squared	0.708629		S.D. dependent var	0.220345
S.E. of regression	0.118940		Akaike info criterion	-1.389775
Sum squared resid	32.15542		F-statistic	80.21044
Log likelihood	1703.206		Prob(F-statistic)	0.000000

LL_IMPDIST: doubly log-transformed implied distance; *C*: constant; *L_IMPDISTSF*: estimated sums of implied distances per region; *RDXDIFF*: angle (rad) between research expenditure vectors (on the business, government and education sector, respectively); *SQRT_GEODIST*: square root of geographic great circle distance between regional centres of gravity; *INTRAROMANIC*: dummy variable with the value of 1 for observations denoting a bilateral connection between regions speaking a Romanic language or Greek. Regional dummy variables (68) were included, but are not displayed.

Table A.11: Estimation results for implied distance modeling with bilateral factors and regional dummies

Dependent Variable: LL_IMPDIST
Method: Least Squares
Included observations: 2278

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.547206	0.057257	9.557083	0.0000
RDXDIFF	0.162201	0.019589	8.280040	0.0000
SQRT_GEODIST	0.004968	0.000360	13.81406	0.0000
INTRAROMANIC	-0.078640	0.011034	-7.127100	0.0000
DUMMY_AT1	-0.020010	0.047232	-0.423646	0.6719
DUMMY_AT2	0.070858	0.044548	1.590594	0.1118
DUMMY_AT3	0.043463	0.043287	1.004064	0.3155
DUMMY_BE1	0.141307	0.040764	3.466485	0.0005
DUMMY_BE2	-0.144693	0.039614	-3.652576	0.0003
DUMMY_BE3	0.132513	0.038090	3.478976	0.0005
DUMMY_DE1	-0.116512	0.037409	-3.114532	0.0019
DUMMY_DE2	0.041310	0.036675	1.126381	0.2601
DUMMY_DE3	-0.138431	0.036145	-3.829859	0.0001
DUMMY_DE4	0.067324	0.035847	1.878073	0.0605
DUMMY_DE5	-0.123992	0.035217	-3.520800	0.0004
DUMMY_DE6	0.269980	0.034960	7.722635	0.0000
DUMMY_DE7	-0.300502	0.034751	-8.647305	0.0000
DUMMY_DE8	0.279368	0.034589	8.076776	0.0000
DUMMY_DE9	-0.014331	0.034515	-0.415219	0.6780
DUMMY_DEA	0.002334	0.034186	0.068272	0.9456
DUMMY_DEB	-0.057405	0.034280	-1.674576	0.0942
DUMMY_DEC	0.190178	0.034050	5.585331	0.0000
DUMMY_DED	-0.054766	0.034170	-1.602733	0.1091
DUMMY_DEE	0.212032	0.034086	6.220447	0.0000
DUMMY_DEF	0.085265	0.034089	2.501252	0.0124
DUMMY_DEG	0.137054	0.034025	4.028070	0.0001
DUMMY_DK0	0.168745	0.034068	4.953143	0.0000
DUMMY_ES1	-0.186360	0.034117	-5.462417	0.0000
DUMMY_ES2	0.341369	0.034395	9.924854	0.0000
DUMMY_ES3	-0.122029	0.034020	-3.586940	0.0003
DUMMY_ES4	0.084747	0.034496	2.456736	0.0141
DUMMY_ES5	0.169172	0.034003	4.975151	0.0000
DUMMY_ES6	-0.128341	0.034191	-3.753605	0.0002
DUMMY_ES7	0.183828	0.034302	5.359180	0.0000
DUMMY_FI1	0.027769	0.035386	0.784750	0.4327
DUMMY_FR1	-0.125270	0.034247	-3.657847	0.0003
DUMMY_FR2	0.038016	0.034212	1.111189	0.2666
DUMMY_FR3	0.037766	0.034132	1.106462	0.2686
DUMMY_FR4	0.096351	0.034005	2.833428	0.0046
DUMMY_FR5	0.030520	0.034139	0.893992	0.3714
DUMMY_FR6	0.038021	0.034242	1.110369	0.2670

DUMMY_FR7	0.019154	0.034181	0.560356	0.5753
DUMMY_FR8	0.008085	0.034120	0.236952	0.8127
DUMMY_GR1	0.053381	0.034324	1.555179	0.1200
DUMMY_GR2	0.033691	0.035399	0.951740	0.3413
DUMMY_GR3	0.011597	0.034622	0.334967	0.7377
DUMMY_GR4	-0.098643	0.035666	-2.765723	0.0057
DUMMY_IE0	0.055281	0.034392	1.607415	0.1081
DUMMY_ITC	-0.111573	0.033877	-3.293505	0.0010
DUMMY_ITD	0.051860	0.034119	1.519966	0.1287
DUMMY_ITE	-0.046731	0.033876	-1.379474	0.1679
DUMMY_ITF	0.010629	0.034211	0.310689	0.7561
DUMMY_ITG	0.014679	0.034097	0.430508	0.6669
DUMMY_LU0	0.139597	0.034564	4.038767	0.0001
DUMMY_NL1	0.239978	0.033200	7.228265	0.0000
DUMMY_NL2	-0.045433	0.034553	-1.314874	0.1887
DUMMY_NL3	0.007605	0.032905	0.231112	0.8172
DUMMY_NL4	-0.092829	0.034702	-2.675057	0.0075
DUMMY_PT1	0.177304	0.032431	5.467198	0.0000
DUMMY_SE0	-0.249796	0.035703	-6.996546	0.0000
DUMMY_UKC	0.119661	0.032187	3.717656	0.0002
DUMMY_UKD	-0.045897	0.035028	-1.310297	0.1902
DUMMY_UKE	0.002822	0.031296	0.090180	0.9282
DUMMY_UKF	0.003575	0.034920	0.102390	0.9185
DUMMY_UKG	-0.014429	0.030024	-0.480580	0.6309
DUMMY_UKH	0.062841	0.034810	1.805270	0.0712
DUMMY_UKI	-0.192512	0.028066	-6.859349	0.0000
DUMMY_UKJ	0.100940	0.034380	2.936009	0.0034
DUMMY_UKK	-0.249336	0.024999	-9.974006	0.0000
DUMMY_UKL	0.212042	0.033598	6.311120	0.0000
DUMMY_UKM	-0.183594	0.019470	-9.429767	0.0000
DUMMY_UKN	0.169928	0.031924	5.322816	0.0000

R-squared	0.712627	Mean dependent var	0.838110
Adjusted R-squared	0.703654	S.D. dependent var	0.220345
S.E. of regression	0.119951	Akaike info criterion	-1.373259
Sum squared resid	32.71879	F-statistic	79.42331
Log likelihood	1682.833	Prob(F-statistic)	0.000000

LL_IMPDIST: doubly log-transformed implied distance; *C*: constant; *RDXDIFF*: angle (rad) between research expenditure vectors (on the business, government and education sector, respectively); *SQRT_GEODIST*: square root of geographic great circle distance between regional centres of gravity; *INTRAROMANIC*: dummy variable with the value of 1 for observations denoting a bilateral connection between regions speaking a Romanic language or Greek, and 0 otherwise; *DUMMY_x*: Regional dummy of value 1 if the respective NUTS-1 region is involved into an observation, and 0 otherwise;

Table A.12: Variable evaluation results for the estimation of log-transformed implied masses with prospective variable pairs - selected evaluation results ranked by Adjusted R^2

69 prospective variables = 2346 pairings *68 observations*
Spec: LS L_IMPMASS C Var1 Var2 *30 runs per pair*
Dependent: L_IMPMASS *Fixed Exogenous: Constant C*

1st Variable	2nd Variable	random selection summary results		Entire sample estimation statistics			
Var1	Var2	mean F-Stat	% Prob.	adj.R2	t-stat Const	t-stat Var1	t-stat Var2
L_GDP_MP	*L_RDXPNAT_BIZ*	1.10	0.93	0.794	-2.95	11.32	8.31
L_GDP_MP	*L_RDXPNAT_TOT*	1.04	0.97	0.793	-3.09	11.47	8.29
L_GDP_MP	*L_RSTAPNAT_BIZ*	0.98	1.00	0.792	-3.01	11.36	8.26
L_GDP_MP	*L_RDXPNAT_GOV*	1.17	0.97	0.789	-3.34	11.74	8.14
L_GDP_MP	*L_RSTAPNAT_TOT*	1.05	1.00	0.789	-3.19	11.51	8.13
L_GDP_MP	*L_RSTAPNAT_GOV*	1.05	1.00	0.787	-3.41	11.78	8.05
L_KIS_HRST	*L_RDXPNAT_BIZ*	1.13	1.00	0.785	6.09	10.99	8.16
L_KIS_HRST	*L_RDXPNAT_TOT*	1.22	0.93	0.784	5.94	11.11	8.13
L_KIS_HRST	*L_RSTAPNAT_BIZ*	0.91	1.00	0.784	6.01	11.03	8.11
L_KIS_HRSTC	*L_RDXPNAT_BIZ*	1.22	0.93	0.782	7.62	10.85	7.90
L_NBBIZ25	*L_RDXPNAT_TOT*	0.92	0.97	0.781	10.25	10.98	8.75
L_KIS_HRSTC	*L_RSTAPNAT_BIZ*	0.91	1.00	0.780	7.54	10.88	7.84
L_KIS_HRST	*L_RSTAPNAT_TOT*	1.13	1.00	0.780	5.77	11.15	7.97
L_NBBIZ25	*L_RSTAPNAT_TOT*	1.24	1.00	0.779	10.05	11.12	8.68
L_KIS_HRSTE	*L_RDXPNAT_BIZ*	1.05	1.00	0.779	7.50	10.75	7.95
L_KIS_HRSTC	*L_RDXPNAT_TOT*	1.08	1.00	0.779	7.46	10.91	7.80
L_KIS_HRSTO	*L_RDXPNAT_BIZ*	1.09	0.97	0.778	6.31	10.69	8.05
L_KIS_HRSTE	*L_RSTAPNAT_BIZ*	1.09	1.00	0.778	7.41	10.78	7.89
L_KIS_HRSTE	*L_RDXPNAT_TOT*	1.08	1.00	0.777	7.33	10.82	7.86
L_KIS_HRSTO	*L_RDXPNAT_TOT*	1.05	1.00	0.776	6.16	10.81	8.00
L_KIS_HRSTO	*L_RSTAPNAT_BIZ*	1.13	1.00	0.776	6.23	10.73	8.00
L_NBBIZ25	*L_RDXPNAT_GOV*	1.15	1.00	0.776	9.95	11.22	8.57
L_KIS_HRST	*L_RDXPNAT_GOV*	1.03	1.00	0.776	5.68	11.21	7.81
L_NBBIZ25	*L_RSTAPNAT_GOV*	1.16	1.00	0.775	9.82	11.31	8.53
L_KIS_HRSTC	*L_RSTAPNAT_TOT*	0.94	1.00	0.774	7.27	10.92	7.62
L_KIS_HRST	*L_RSTAPNAT_GOV*	0.94	1.00	0.772	5.57	11.20	7.68
L_KIS_HRSTO	*L_RSTAPNAT_TOT*	1.08	1.00	0.772	5.99	10.85	7.84
L_KIS_HRSTE	*L_RSTAPNAT_TOT*	1.11	1.00	0.772	7.15	10.84	7.69
L_TOPBIZ	*L_RDXPNAT_BIZ*	0.95	1.00	0.770	2.06	10.41	7.50
L_KIS_HRSTC	*L_RDXPNAT_GOV*	1.05	0.97	0.770	7.18	10.99	7.47
L_TOPBIZ	*L_RDXPNAT_TOT*	0.98	1.00	0.770	1.93	10.56	7.49
L_NBBIZ25	*L_RDXPNAT_BIZ*	1.14	1.00	0.770	10.30	10.40	8.34
L_KIS_HRSTO	*L_RDXPNAT_GOV*	0.90	1.00	0.769	5.90	10.96	7.74
L_TOPBIZ	*L_RSTAPNAT_BIZ*	1.00	1.00	0.768	2.00	10.44	7.44
L_GDP_MP	*L_RDXPNAT_EDU*	1.08	1.00	0.768	-3.29	11.28	7.37
L_NBBIZ25	*L_RSTAPNAT_BIZ*	1.02	0.97	0.768	10.19	10.41	8.26
L_GDP_MP	*L_RSTAPNAT_EDU*	1.04	1.00	0.767	-3.37	11.36	7.34
L_TOPBIZ	*L_RSTAPNAT_TOT*	1.09	1.00	0.767	1.79	10.65	7.38
L_KIS_HRSTE	*L_RDXPNAT_GOV*	1.10	1.00	0.766	7.05	10.85	7.48
L_KIS_HRSTC	*L_RSTAPNAT_GOV*	1.14	0.97	0.765	7.06	10.95	7.31
L_GQ_HRST	*L_RDXPNAT_BIZ*	0.97	1.00	0.765	4.67	10.23	8.29
L_KIS_HRSTO	*L_RSTAPNAT_GOV*	1.19	1.00	0.765	5.79	10.94	7.60
L_GQ_HRSTC	*L_RDXPNAT_BIZ*	1.03	1.00	0.765	6.54	10.23	7.90
L_GQ_HRST	*L_RDXPNAT_TOT*	1.03	1.00	0.764	4.51	10.36	8.26
L_TOPBIZ	*L_RDXPNAT_GOV*	0.94	1.00	0.764	1.71	10.77	7.29
L_GQ_HRST	*L_RSTAPNAT_BIZ*	0.96	1.00	0.763	4.59	10.26	8.23

L_GQ_HRSTC	*L_RSTAPNAT_BIZ*	1.07	1.00	0.763	6.46	10.25	7.83
L_NBBIZ25	*L_RDXPNAT_EDU*	1.09	0.97	0.763	9.42	11.08	8.09
L_TOPBIZ	*L_RSTAPNAT_GOV*	1.06	1.00	0.762	1.62	10.85	7.24
L_RSTAFF_TOT	*L_RDXPNAT_TOT*	1.07	1.00	0.762	0.86	10.29	6.78
L_NBBIZ25	*L_RSTAPNAT_EDU*	1.03	1.00	0.762	9.30	11.18	8.08
L_GQ_HRSTC	*L_RDXPNAT_TOT*	0.97	1.00	0.762	6.39	10.29	7.80
L_RSTAFF_TOT	*L_RSTAPNAT_BIZ*	1.05	1.00	0.762	0.93	10.20	6.75
L_KIS_HRSTE	*L_RSTAPNAT_GOV*	1.01	1.00	0.762	6.93	10.82	7.33
L_RSTAFF_TOT	*L_RDXPNAT_BIZ*	1.10	1.00	0.760	1.00	10.07	6.71
…	…	…	…	…	…	…	…
L_RDX_MPGOV	*L_RDXPNAT_GOV*	1.25	0.93	0.535	16.35	5.20	5.40
L_TOT_HRST	*XPRES_BIZ*	1.04	1.00	0.535	1.55	7.42	2.71
L_RDX_MPTOT3	*RSTAFF_PCLBIZ*	1.05	0.97	0.535	4.02	7.57	-1.71
L_NBBIZ25	*RSTAFF_PCLBIZ*	1.03	1.00	0.535	7.01	7.57	1.27
L_TOT_HRST	*L_GQ_HRST*	1.07	1.00	0.535	3.13	-1.59	2.70
L_GQ_HRST	*TOPFIRMSPC*	1.08	0.97	0.535	2.11	8.88	1.59
L_GQ_HRST	*RDXPC_EDU*	1.07	1.00	0.535	3.14	7.57	-1.58
L_RSTAFF_GOV	*L_RSTAPNAT_GOV*	1.11	0.97	0.535	8.05	5.32	4.96
L_TOT_HRSTC	*L_RSTAFF_EDU*	1.08	1.00	0.534	2.48	4.19	1.58
L_RDX_MPGOV	*L_NBBIZ10*	0.99	1.00	0.534	13.76	1.75	5.38
L_GQ_HRST	*L_EMP_PFIRM*	1.04	1.00	0.534	1.98	7.66	1.56
L_TOT_HRSTC	*RDXPC_BIZ*	1.04	1.00	0.534	4.06	7.78	1.57
L_GQ_HRST	*RDXPC_BIZ*	0.97	1.00	0.534	2.66	7.78	1.55
L_RSTAFF_GOV	*L_PATPRX_BIZ*	1.08	1.00	0.534	8.26	7.56	-4.95
L_TOT_HRSTC	*L_RSTAFF_GOV*	0.94	1.00	0.534	3.74	4.95	1.56
L_EMPLOYEES	*L_RDX_MPBIZ*	1.09	1.00	0.534	1.93	3.43	2.65
L_POP	*L_RDX_PCYBIZ*	1.08	1.00	0.534	-0.07	7.27	3.78
L_TOT_HRST	*L_PAT_TOPBIZ*	1.12	1.00	0.533	1.94	8.87	-2.66
L_GQ_HRST	*L_RDX_PCYBIZ*	1.13	1.00	0.533	2.97	7.26	1.51
L_EMPLOYEES	*L_NBBIZ25*	1.09	1.00	0.533	2.25	1.16	2.63
L_TOT_HRSTC	*RDXPC_EDU*	1.05	1.00	0.533	4.32	7.54	-1.51
L_GQ_HRST	*L_RDX_MEBIZ*	1.03	0.97	0.533	3.07	4.46	1.48
L_NBBIZ10	*RSTAFF_PCLGOV*	1.07	0.93	0.532	16.10	8.44	1.68
L_GQ_HRST	*L_NBBIZ*	1.15	1.00	0.532	1.50	3.59	1.47
L_GQ_HRST	*RSTAFF_PCLGOV*	0.99	0.97	0.532	2.70	8.43	1.46
…	…	…	…	…	…	…	…
ENGLISH	*L_PAT_TOPBIZ*	0.89	1.00	-0.026	14.81	-0.56	0.04
ENGLISH	*XPRES_GOV*	1.05	1.00	-0.026	13.72	-0.55	0.03
L_RDX_PCYEDU	*XPRES_GOV*	0.92	1.00	-0.027	15.74	0.46	-0.21
L_RDX_PCYEDU	*TOPFIRMSPC*	1.16	0.97	-0.027	20.36	0.38	-0.18
XPRES_GOV	*TOP10_ALLBIZ*	1.08	1.00	-0.028	18.20	-0.26	0.41
L_RDX_PCYEDU	*L_PAT_TOPBIZ*	1.08	1.00	-0.028	21.63	0.42	0.00
TOP10_ALLBIZ	*L_PAT_TOPBIZ*	1.15	1.00	-0.029	38.78	0.34	0.03
TOPFIRMSPC	*L_PAT_TOPBIZ*	1.30	0.90	-0.029	38.90	-0.30	0.16
XPRES_GOV	*TOPFIRMSPC*	1.22	0.97	-0.030	18.39	-0.07	-0.24
XPRES_GOV	*L_PAT_TOPBIZ*	1.12	1.00	-0.030	18.38	-0.15	0.12

Notes: Each pair was fitted to the entire sample of 68 observations. Adjusted R^2 and t-statistics are provided under “Entire sample estimation statistics”. For each pair the specification was fitted to two mutually excluding, random sub-samples of 34 observations and a Chow forecast test was performed to test for the consistency of sub-sample estimation results. For each pair this procedure was repeated 30 times. Under “Random selection summary results”, “mean F-Stat” denotes the mean of F-statistics from the Chow forecast test and “% Prob.” provides the share of the draws in which the corresponding F-statistics held a prob value below 0.05. For further details on the evaluation procedure refer to section 6.1.4.

Table A.13: Variable evaluation results for the estimation of log-transformed inter-regional collaborative links with prospective variable pairs - selected evaluation results ranked by R^2 and Chow forecast test performance

67 prospective variables = 2211 pairings	*2346 observations*
Spec: LS L_COOPFP4 C Var1 Var2	*50 runs per pair*
Dependent: L_COOPFP4	*Fixed Exogenous: Constant C*

1st Variable	2nd Variable	mean F-stat	% Prob.	mean coef / stdev(coef)			% Coef significance			R^2 for entire sample			
Var1	Var2			C	Var1	Var2	C	Var1	Var2	OLS R2 w	WLS R2 w	OLS R2 w/o	OLS R2 w/o
L_TOT_HRSTC	*L_TOP5_NB5BIZ*	0.99	0.97	-41.8	69.4	14.0	-1	1	1	0.75	0.71	0.55	0.49
L_RSTAFF_TOT	*XPRES_GOV*	1.04	0.87	15.6	101.6	-18.6	1	1	-1	0.76	0.73	0.63	0.59
L_KIS_HRSTC	*L_TOPBIZ*	0.92	0.90	-37.3	34.1	18.5	-1	1	1	0.77	0.74	0.59	0.54
L_GDP_MP	*RSTAFF_PCLTOT*	0.99	0.83	-43.9	52.8	39.2	-1	1	1	0.77	0.74	0.66	0.61
L_RDX_MPTOT3	*PUBNAT_RST*	0.98	0.80	-57.4	91.0	53.4	-1	1	1	0.76	0.73	0.68	0.64
L_KIS_HRST	*RSTAFF_PCLGOV*	0.98	0.87	-63.5	106.1	24.7	-1	1	1	0.75	0.71	0.61	0.56
L_RDX_MPTOT3	*L_TOPBIZ*	0.99	0.87	-36.1	32.1	21.6	-1	1	1	0.77	0.74	0.61	0.55
L_RSTAFF_TOT	*PUBNAT_RST*	1.19	0.73	-64.6	82.6	29.7	-1	1	1	0.77	0.74	0.73	0.69
L_NBBIZ	*L_RDXPNAT_EDU*	0.97	0.87	-32.3	47.2	29.7	-1	1	1	0.77	0.73	0.59	0.53
L_RDX_MPEDU	*L_RSTAFF_TOT*	1.04	0.83	-53.0	4.2	56.3	-1	0.3	1	0.75	0.71	0.62	0.57
L_RSTAFF_TOT	*SQRT_GEODIST*	1.04	0.83	-45.9	71.4	-5.3	-1	1	-1	0.75	0.71	0.62	0.57
L_KIS_HRST	*L_RDXPNAT_TOT*	1.04	0.77	-48.8	79.8	35.1	-1	1	1	0.77	0.74	0.66	0.61
L_RSTAFF_TOT	*L_TOP5_NB5BIZ*	1.04	0.77	-78.8	105.5	20.9	-1	1	1	0.75	0.72	0.66	0.61
L_RDX_MPGOV	*L_TOPBIZ*	1.04	0.87	-39.8	24.1	43.2	-1	1	1	0.77	0.74	0.56	0.50
L_KIS_HRST	*L_RSTAFF_EDU*	1.06	0.83	-36.0	41.4	5.4	-1	1	0.97	0.75	0.71	0.59	0.53
L_KIS_HRST	*L_GDPP_EMP*	1.04	0.77	-43.7	64.2	27.3	-1	1	1	0.79	0.76	0.65	0.61
L_RDX_PCYBIZ	*L_TOPBIZ*	0.99	0.83	-39.2	31.5	57.4	-1	1	1	0.77	0.74	0.59	0.53
L_RSTAFF_TOT	*L_TOPFIRMSPC*	1.05	0.80	-41.3	56.8	3.9	-1	1	0.03	0.75	0.71	0.62	0.57
L_GDP_MP	*RSTAFF_PCLEDU*	1.00	0.80	-58.5	73.4	17.2	-1	1	1	0.77	0.74	0.61	0.56
L_GDP_MP	*L_RDX_MPEDU*	1.06	0.83	-49.7	49.6	11.2	-1	1	1	0.77	0.73	0.58	0.52
L_KIS_HRSTO	*L_RDXPNAT_EDU*	1.05	0.77	-34.8	57.6	28.8	-1	1	1	0.77	0.74	0.64	0.59
L_RSTAFF_GOV	*L_DRXMP_BG*	1.00	0.80	-49.8	80.6	57.5	-1	1	1	0.76	0.72	0.61	0.56
L_GDP_MP	*CULTDIM*	1.02	0.83	-56.0	79.9	5.2	-1	1	0.9	0.76	0.73	0.58	0.52
L_GDP_MP	*LANGSTUD*	1.08	0.83	-47.4	60.0	3.1	-1	1	0.13	0.76	0.73	0.58	0.52
L_KIS_HRSTO	*L_GDPP_EMP*	1.05	0.77	-40.6	84.4	26.0	-1	1	1	0.79	0.76	0.64	0.59
L_KIS_HRSTC	*L_NBBIZ10*	1.04	0.77	-33.5	33.9	32.4	-1	1	1	0.75	0.72	0.64	0.59
L_RSTAFF_TOT	*L_RDXPNAT_BIZ*	1.05	0.73	-56.8	78.7	23.8	-1	1	1	0.77	0.74	0.67	0.62
L_GDPP_INH	*L_KIS_HRSTC*	1.05	0.73	-44.6	34.3	83.1	-1	1	1	0.76	0.73	0.67	0.62
L_KIS_HRSTE	*RSTAFF_PCLGOV*	1.01	0.80	-51.2	90.8	22.5	-1	1	1	0.75	0.71	0.60	0.54
L_KIS_HRSTC	*L_RDXPNAT_EDU*	1.09	0.77	-35.5	64.5	22.2	-1	1	1	0.77	0.74	0.63	0.58
L_RDX_PCYTOT	*L_RSTAFF_TOT*	1.09	0.77	-49.4	-11.0	61.7	-1	-1	1	0.76	0.72	0.63	0.58
L_RSTAFF_EDU	*L_RDXPNAT_BIZ*	1.01	0.83	-40.6	63.9	42.2	-1	1	1	0.77	0.74	0.56	0.50
L_RSTAFF_TOT	*LANGSTUD*	1.09	0.77	-42.2	63.9	2.7	-1	1	0.2	0.75	0.71	0.63	0.58
L_RSTAFF_TOT	*PATSTRUCT*	1.13	0.77	-47.3	86.0	-12.4	-1	1	-1	0.75	0.71	0.62	0.58
L_RSTAFF_TOT	*XPRES_TOT*	1.09	0.77	3.9	64.4	-5.0	0.8	1	-0.9	0.75	0.71	0.62	0.57
L_GDPP_INH	*L_KIS_HRSTE*	1.03	0.73	-31.0	24.7	70.7	-1	1	1	0.76	0.72	0.65	0.61
L_KIS_HRSTO	*L_RDXPNAT_TOT*	1.06	0.73	-37.4	63.7	32.9	-1	1	1	0.77	0.74	0.65	0.61
L_GDP_MP	*L_RSTAFF_GOV*	1.08	0.80	-48.1	44.1	16.5	-1	1	1	0.76	0.73	0.59	0.53
…	…	…	…	…	…	…	…	…	…	…	…	…	…
L_EMPLOYEES	*L_PAT_TOPBIZ*	1.12	0.67	-38.2	58.3	-4.1	-1	1	-0.33	0.76	0.72	0.48	0.42
L_KIS_HRSTO	*L_TOPFIRMSPC*	1.15	0.57	-37.3	61.4	20.2	-1	1	1	0.75	0.72	0.58	0.53
L_RDX_MPBIZ	*L_RDXDIFF*	1.13	0.67	-8.2	52.1	-9.6	-1	1	-1	0.75	0.72	0.48	0.42

L_RDX_PCYTOT	*L_PATPRX_BIZ*	1.01	0.87	54.8	30.3	-25.6	1	1	-1	0.75	0.72	0.28	0.19
L_NBBIZ	*RSTAFF_PCLTOT*	1.16	0.57	-36.4	52.7	48.1	-1	1	1	0.74	0.71	0.58	0.53
L_RDX_MPGOV	*LANGSTUD*	1.09	0.77	-6.5	48.5	4.3	-1	1	0.53	0.74	0.70	0.38	0.30
L_RDX_MPEDU	*RSTAFF_PCLTOT*	1.11	0.67	-12.3	37.4	29.1	-1	1	1	0.74	0.70	0.48	0.41
L_EMP_PFIRM	*L_RDXPNAT_EDU*	0.96	0.83	-6.6	40.2	24.9	-0.9	1	1	0.76	0.72	0.31	0.23
L_RDX_MPGOV	*SQRT_GEODIST*	1.09	0.77	0.6	43.8	-10.3	0	1	-1	0.74	0.71	0.38	0.30
L_RSTAFF_GOV	*L_RSTAFFDIFF*	1.12	0.70	-15.4	47.2	-14.8	-1	1	-1	0.75	0.71	0.45	0.37
L_RSTAFF_BIZ	*RDXPC_EDU*	1.11	0.67	-21.5	51.2	13.1	-1	1	1	0.74	0.70	0.48	0.41
L_NBBIZ	*RSTAFF_PCLEDU*	1.12	0.63	-41.6	59.4	24.7	-1	1	1	0.74	0.71	0.51	0.45
L_RSTAFF_GOV	*L_EMP_PFIRM*	1.17	0.63	-39.7	44.0	29.9	-1	1	1	0.74	0.70	0.51	0.45
L_EMPLOYEES	*L_RDXDIFF*	1.07	0.63	-41.2	68.6	-18.5	-1	1	-1	0.76	0.72	0.51	0.45
L_TOPBIZ2	*L_DIST2CORE*	1.15	0.67	-15.4	38.2	-23.1	-1	1	-1	0.77	0.74	0.48	0.41
…	…	…	…	…	…	…	…	…	…	…	…	…	…
L_TOP10_ABIZ	*L_GDPP_EMP*	1.36	0.40	-11.0	1.1	17.8	-1	0	1	0.78	0.74	0.07	-0.05
L_DRXMP_BG	*L_PAT_TOPBIZ*	1.21	0.43	21.1	9.5	0.9	1	1	0.07	0.77	0.73	0.04	-0.08
L_PUBNAT_POP	*L_RDX_PTOP2_*	1.33	0.33	-14.3	20.6	14.1	-1	1	1	0.74	0.71	0.13	0.02
XPRES_TOT	*L_TOPFIRMSPC*	1.36	0.33	-25.8	26.2	-10.3	-1	1	-0.9	0.74	0.70	0.12	0.01
L_GDPP_INH	*L_RDXPNAT_GOV*	1.63	0.20	-18.1	21.8	14.8	-1	1	1	0.76	0.72	0.25	0.14
L_PUBNAT_POP	*L_TOP5_NB5BIZ*	1.26	0.37	-14.5	21.5	4.9	-1	1	0.47	0.74	0.71	0.08	-0.03
L_RDX_PCYGOV	*XPRES_EDU*	1.29	0.33	-10.4	11.3	10.6	-1	1	1	0.75	0.71	0.07	-0.05
L_TOP10_ABIZ	*L_RDX_PTOP2_*	1.29	0.33	86.3	0.0	16.1	1	0	1	0.74	0.70	0.06	-0.05
L_PATPST_BIZ	*L_PAT_TOPBIZ*	1.28	0.33	19.2	-15.0	13.2	1	-1	1	0.75	0.72	0.06	-0.05
L_PAT_MLF	*L_TOPFIRMSPC*	1.59	0.37	26.6	0.0	-3.9	1	0	0	0.74	0.71	0.00	-0.12

Notes: Each pair was fitted to the entire sample of 2346 observations. For each pair the specification was fitted to two mutually excluding, random sub-samples of 34 (34+1) /2 = 595 observations and a Chow forecast test was performed to test for the consistency of sub-sample estimation results. For each pair this procedure was repeated 50 times. "Mean F-Stat" denotes the mean of F-statistics from the Chow forecast test and "% Prob." provides the share of the draws in which the corresponding F-statistics held a prob value below 0.05. For further details on the evaluation procedure refer to section 6.1.4.
"mean coef / stdev(coef)" intends to indicate the stability of estimation coefficients over the 50 draws and provides the mean of the coefficient divided by their standard deviation over the 50 draws. "% Coef significance" declares the share of draws in which the respective coefficient's t-statistics indicated a Type I error probability < 0.05. "R2 for entire sample" denotes the R^2 for the structure estimated on the total 2346 observations. "OLS R2 w" indicates the OLS estimation's R2 when *L_IMPDISTF*, the fit from Estimation 3 was included as an additional regressor (in order to evaluate factors against a proxy for distance). Conversely, "OLS R2 w/o" indicates the OLS estimation's R2 for the pure structure, i.e. *L_COOPFP4*, a constant and the two prospective explanatory variables. "WLS" refers the estimation results when the estimation was weighted to the following factor: A multiplicative of the squared sum of (68 each) errors for region i and the squared sum of (68 each) errors for region j.

Table A.14: Example printout for consistency check procedure on the structure in Estimation 7 with 25 draws

F-stat	Prob F-stat	Adj.R2 subset 1	subset 1 nodes	Coefficients								t-statistics							
				1)	2)	3)	4)	5)	6)	7)	8)	1)	2)	3)	4)	5)	6)	7)	8)
0.72	1.00	0.86	AT1 AT3 BE3 ...	16.8	0.17	0.16	-1.34	-0.57	0.53	-0.50	-1.31	16.4	7.5	13.8	-10.1	-7.4	12.1	-9.9	-21.7
0.74	1.00	0.85	AT2 AT3 BE1 ...	16.8	0.17	0.14	-1.30	-0.57	0.62	-0.46	-1.33	16.3	7.6	12.4	-9.9	-7.5	14.1	-9.1	-21.7
0.79	1.00	0.85	AT1 AT2 AT3 ...	16.8	0.16	0.16	-1.33	-0.59	0.57	-0.49	-1.30	16.3	7.2	13.8	-10.1	-7.7	13.3	-9.7	-21.3
0.81	0.99	0.85	AT2 AT3 BE1 ...	17.3	0.16	0.14	-1.44	-0.55	0.62	-0.50	-1.36	16.9	7.1	11.9	-11.1	-7.2	14.1	-10.0	-22.1
0.82	0.99	0.85	AT1 AT2 BE1 ...	16.7	0.17	0.16	-1.18	-0.61	0.56	-0.48	-1.30	16.5	7.5	13.9	-9.3	-8.0	12.8	-9.8	-21.8
0.83	0.99	0.85	AT1 AT3 BE1 ...	17.0	0.16	0.17	-1.20	-0.63	0.55	-0.51	-1.30	16.6	6.9	14.7	-9.2	-8.3	12.5	-10.1	-21.2
0.85	0.98	0.85	AT1 AT3 BE2 ...	16.3	0.18	0.16	-1.10	-0.64	0.57	-0.45	-1.28	16.2	8.1	14.1	-8.5	-8.3	12.9	-9.0	-21.3
0.88	0.95	0.85	AT2 AT3 BE2 ...	17.2	0.15	0.15	-1.27	-0.50	0.59	-0.54	-1.33	16.5	6.7	13.3	-9.6	-6.4	13.3	-10.7	-21.7
0.90	0.89	0.85	AT2 AT3 BE1 ...	16.5	0.17	0.14	-1.23	-0.56	0.57	-0.56	-1.30	16.0	7.7	12.1	-9.6	-7.3	13.0	-11.1	-21.3
0.96	0.70	0.84	AT1 AT3 BE1 ...	16.4	0.16	0.17	-1.19	-0.67	0.56	-0.47	-1.25	15.6	6.9	14.0	-8.9	-8.4	12.5	-9.1	-19.9
0.96	0.68	0.85	BE1 BE2 DE2 ...	17.9	0.15	0.14	-1.31	-0.60	0.56	-0.45	-1.38	17.3	6.7	11.7	-9.9	-7.6	12.8	-8.9	-22.6
0.98	0.58	0.86	AT1 BE2 DE1 ...	15.5	0.19	0.16	-1.25	-0.53	0.65	-0.40	-1.24	15.6	8.9	14.9	-9.8	-7.3	15.3	-8.0	-21.0
1.04	0.30	0.85	AT3 BE1 BE3 ...	16.1	0.18	0.14	-1.24	-0.55	0.57	-0.49	-1.27	15.9	8.1	13.0	-9.8	-7.5	13.2	-10.0	-21.2
1.08	0.17	0.85	AT1 AT2 BE1 ...	17.9	0.15	0.15	-1.13	-0.62	0.61	-0.52	-1.38	17.2	6.6	12.9	-8.4	-7.9	13.6	-10.0	-22.2
1.12	0.08	0.85	AT2 DE1 DE2 ...	16.7	0.17	0.15	-1.24	-0.56	0.57	-0.50	-1.31	16.3	7.4	13.1	-9.7	-7.1	13.1	-10.1	-21.5
1.15	0.04	0.85	AT2 AT3 BE1 ...	17.2	0.16	0.15	-1.25	-0.62	0.65	-0.42	-1.34	17.0	7.3	13.1	-9.8	-8.3	15.4	-8.5	-22.2
1.17	0.03	0.85	AT2 BE1 BE2 ...	17.2	0.16	0.15	-1.35	-0.54	0.60	-0.49	-1.33	17.1	7.2	13.5	-10.7	-7.3	13.8	-9.9	-22.4
1.21	0.01	0.85	AT2 DE2 DE3 ...	16.9	0.16	0.15	-1.35	-0.56	0.57	-0.52	-1.31	16.3	7.1	13.5	-10.4	-7.1	13.0	-10.3	-21.3
1.24	0.00	0.85	AT2 AT3 BE1 ...	16.4	0.17	0.17	-1.22	-0.58	0.58	-0.48	-1.27	16.5	7.7	15.1	-9.6	-7.8	13.6	-9.6	-21.5
1.27	0.00	0.85	AT2 AT3 BE1 ...	17.5	0.15	0.14	-1.21	-0.60	0.57	-0.51	-1.35	17.1	6.6	12.4	-9.1	-7.9	12.9	-10.0	-22.1
1.32	0.00	0.84	AT2 BE1 BE3 ...	16.2	0.18	0.14	-1.28	-0.53	0.58	-0.50	-1.29	15.7	7.9	12.3	-9.5	-6.8	12.8	-9.9	-20.8
1.39	0.00	0.85	AT1 BE1 DE1 ...	15.0	0.20	0.15	-1.19	-0.49	0.55	-0.51	-1.22	15.0	9.1	13.3	-9.4	-6.6	13.0	-10.4	-20.5
1.40	0.00	0.85	BE1 BE2 DE4 ...	15.7	0.19	0.16	-1.25	-0.64	0.55	-0.52	-1.24	15.7	8.4	13.8	-9.5	-8.3	12.6	-10.4	-20.9
1.48	0.00	0.85	AT3 BE1 BE3 ...	16.8	0.18	0.14	-1.22	-0.64	0.64	-0.59	-1.33	16.2	7.9	11.8	-9.3	-8.4	14.3	-11.8	-21.6
1.56	0.00	0.85	AT3 BE3 DE1 ...	17.0	0.16	0.15	-1.18	-0.64	0.59	-0.46	-1.31	16.7	7.3	13.2	-9.0	-8.3	13.5	-9.3	-21.7

Variables in structure 1) *C* 2) *L_RSTAFF_TOT* 3) *LL_RDXPNAT_T* 4) *L_RDXDIFF* 5) *PATSTRUCT* 6) *INTRAROMANIC* 7) *DUMMY_DDR* 8) *L_IMPDISTSF*

Notes: For 25 times the structure was fitted to two mutually excluding, random sub-samples of 34 (34+1) /2 = 595 observations and a Chow forecast test was performed to test for the consistency of sub-sample estimation results. “F-Stat” denotes the F-statistic from the Chow forecast test and “Prob. F-Stat” its Type I error probability. “Adj. R2” denotes the adjusted R^2 for the one of the first of the two sub-samples. The 34 regions in the first subset are provided in “subset 1 nodes” (not fully displayed). “Coefficents” displays the coefficient values in each estimation, “t-statistics” their t-stats. For further details on the consistency check procedure refer to section 6.1.3.

APPENDIX B: CODE SNIPPETS

The empirical analysis was intended to be done with Stata, but later on we had to switch to EViews due to matrix size limitations in our Stata version. The code written in the course of this diploma thesis encompasses many more than the little EViews programs outlined below. The most compelling tasks were the variable selection and consistency check procedures which are outlined below. Subroutines comprise a routine for calculating weights for WLS. Table B.1 lines out the structure of the subroutines. Moreover, procedures Procedure 9 and Procedure 10 comprise matrix operations used for computations, and Procedure 12, a MS Excel VBA procedure, performs the output scaling method described in section 6.1.5.

Contents of Appendix B

EViews Variable Evaluation and Consistency Check Procedure

The layout of the variable evaluation procedure is depicted in Table B.1: The metafile montecarlo2.prg samples the pairs of variables and calls the consistency check procedure montecarlo.prg. For a pre-defined number of times, the latter procedure in turn calls a program extracting two distinct subsets of random nodes and two smaller procedures in order to compute the Chow forecast test statistic. The following pages present the code for the seven sub-routines needed. The sub-routines may be executed alone, i.e. with being called by a "parent", but all of them require parameters/objects, which may either be provided when running the routine, or which may be set to default values in the code.

Since the routines work with string operators, it is not feasible to include operators in a group member or equation specification (for instance *LS LOG(coopfp4 +1) C LOG(geodist +1)*). Even the operator "C" for constant may cause problems. Instead, please work with a specification only consisting of series labels (e.g.: *l_coopfp4 oneser l_geodist*, where the former and the latter are series containing log-transformed values and oneser is a series with all observations=1 and zero variance)

The parameters/objects needed are the following:

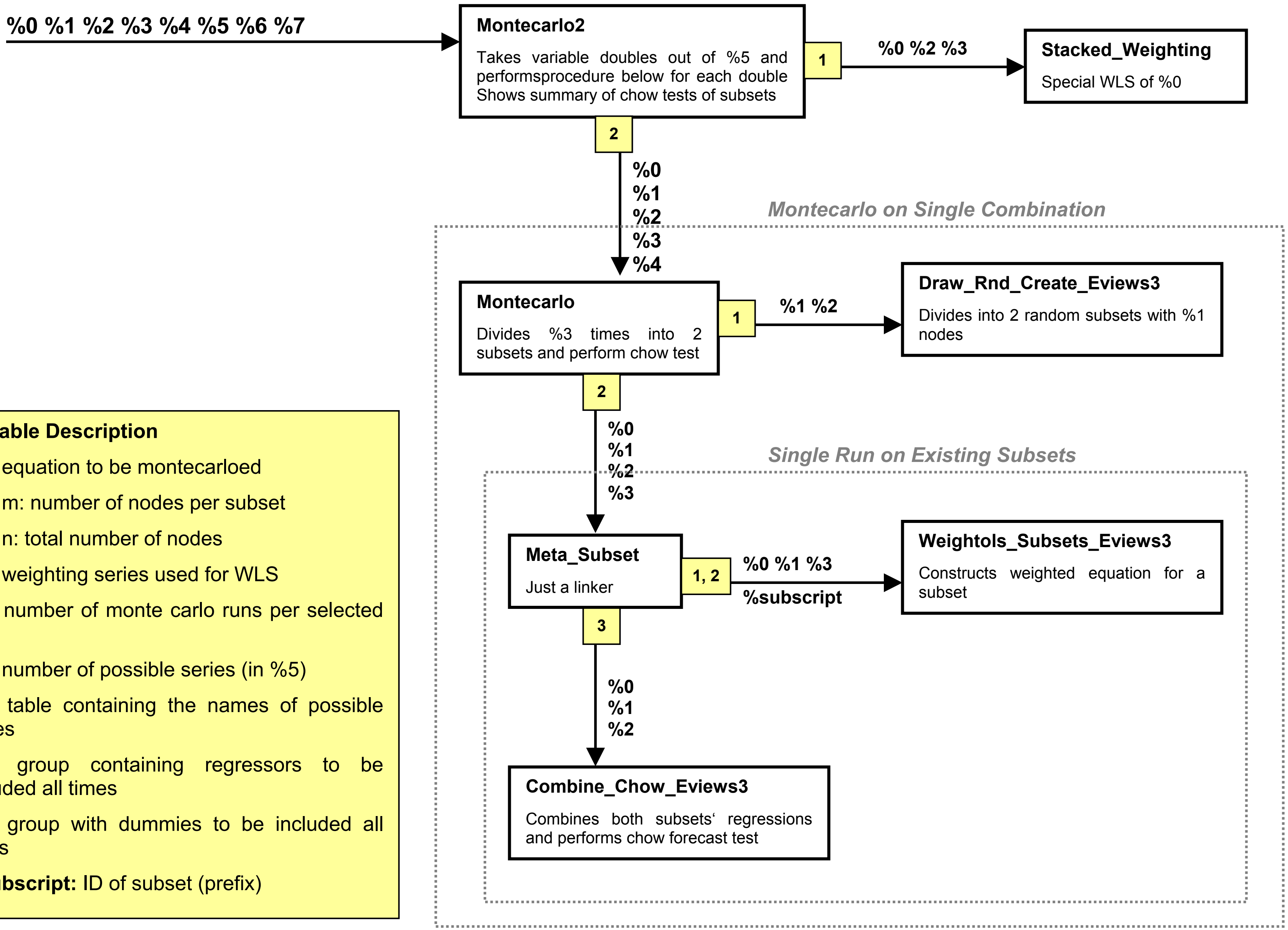
1. **%0:** An Eviews equation object, at best following the structure LS *Dependent RegressorGroup* (see %7) *DummyGroup* (see %8)
2. **%1:** The number of nodes per random subset to be drawn (integer, at best half of %2)
3. **%2:** The integer number of nodes in the interaction pattern in the sample (e.g. this thesis took stacked series of 2346 observations which originated from a 68 x 68 matrix, i.e. 68 nodes)
4. **%3:** A weighting series needed for WLS of %0 (If no WLS is required, take a series with zero variance)
5. **%4:** The integer number of runs, i.e. the number of times the consistency test (fitting a structure to distinct sub-samples and performing the Chow test) is carried out
6. **%5:** The integer number of prospective explanatory series to be evaluated in pairs (their names are stored in %6)
7. **%6:** Table, whose first column contains the labels/identifiers of the prospective explanatory series

8. **%7:** Regressor Group containing the independent series in the equation object %0 (may be empty)
9. **%8:** Dummy Group containing additional independent series in the equation object %0 not included in %7 (may be empty)

For instance: In order to test the entire set of programmes open the EViews workfile on the enclosed CD, and then type the following into the EViews command prompt:

```
run montecarlo2.prg unrstr_eval 34 68 oneser 2 3 unrstr_choice2 unrstr_rgroup unrstr_dgroup
```

Table B. 1: Node-specific WLS, Chow and Iteration Flowchart



Variable Description

%0: equation to be montecarloed

%1: m: number of nodes per subset

%2: n: total number of nodes

%3: weighting series used for WLS

%4: number of monte carlo runs per selected pair

%5: number of possible series (in %5)

%6: table containing the names of possible series

%7: group containing regressors to be included all times

%8: group with dummies to be included all times

%subscript: ID of subset (prefix)

Procedure 1: Montecarlo2.prg (Eviews)

```
'This procedure creates a new workfile with the equation structure of %0 and evaluates all variable pairs given in %6 by calling montecarlo.prg

'EQUATION %0 MUST NOT INCLUDE OPERATORS:
'i.e. no LOG(ser01) or ser01^2 etc.
'Do not use "C" as the constant, use a series containing value 1 instead (e.g. ONESER)

'Objects needed:
'  labels (table): contains at least %2 IDs denoting nodes in 1st column;
'  %0 (equation);
'  %3 (series): weighting series, if no WLS required, give oneser;
'  %6 (table): contains at least %5 series names denoting series to be evaluated in 1st column
'  %7 (group): contains regressors to be included into every evaluation (may be empty)
'  %8 (group): contains dummy regressors to be included into every evaluation (may be empty)

'Variables needed:
'%1: nodes per subset; %2: nodes in total sample; %4: number of runs; %5: number of series to be evaluated;

'=================================
'Setting default values for optional parameters
if %0="" then
  %0="versuch1"                    'equation to be montecarloed
endif
if %1="" then
  %1="34"                   'number of nodes per subset
endif
if %2="" then
  %2="68"                   'n: number of nodes
endif
if %3="" then
  %3="impdist"                     'weighting series
endif
if %4="" then
  %4="20"                   'number of monte carlo runs per selected pair
endif
if %5="" then
  %5="32"                   'number of possible series
endif
if %6="" then
  %6="var_choice"           'table containing the names of possible series in column 1
endif
if %7="" then
  %7="hihi"                        'group containing regressors to be included into all runs
endif
if %8="" then
  %8="hihidummy"            'group containing dummies to be included into all runs
endif

'=============================
'Transform optional parameters to variables and clear the parameters
!n=@val(%2)
%weightingseries=%3
!numberruns=@val(%4)
!novars=@val(%5)
%5=""
%vartable=%6
%6=""
%reggroup=%7
%7=""
%reggroupdummy=%8
%8=""

'Get dependent var
  {%0}.ls
  {%0}.makeregs temp_regs
  %depi=temp_regs.@seriesname(1)
  d temp_regs

'Create a new workfile to speed up process
  dbcreate uebersiedl
  copy labels ::labels
  copy %depi ::{%depi}
  copy %0 ::{%0}
  copy %vartable ::{%vartable}
  copy %reggroup ::{%reggroup}
  copy %reggroupdummy ::{%reggroupdummy}
```

```
 copy %weightingseries ::{%weightingseries}

 for !i=1 to !novars      'Copy series denoted in %vartable
   %curr ={%vartable}(!i,1)
   copy %curr ::%curr
 next !i
 !smpllgth=!n*(!n+1)/2
 create mcruns u 1 !smpllgth
 fetch ::*
 dbdelete uebersiedl

'------------------
'Continue in new workfile

 include stacked_weighting   'Do WLS of %0
 !otherregs={%0}.@ncoef

'Layout result table
 table mc_results
 d mc_results
 table mc_results
 mc_results(1,3)="mean f-stat"
 mc_results(1,4)="% insign p-val"
 mc_results(1,5)="mean coeff / stdev"

'=====================
'GET RESULTS INTO TABLE
'=====================

for !vars1=1 to !novars

        for !vars2=!vars1+1 to !novars

                !crossindex=!vars2-1+(!novars-1-!vars1/2)*(!vars1-1) 'index of this combination in (!novars-1)(!novars-2)/2
combinations
                %current1={%vartable}(!vars1, 1)
                %current2={%vartable}(!vars2, 1)
                %status= "step "+ @str(!crossindex)+" with "+%current1+" "+%current2
                statusline %status

                {%reggroup}.add %current1 %current2 'Add vars to evaluate to regressor group

                {%0}.ls    'Re-fit %0
                'include stacked_weighting.prg %0
                include montecarlo %0 %1 %2 %3 %4    'Call montecarlo.prg
                {%reggroup}.drop %current1 %current2  'Drop both variables to be evaluated
                d _1_*   'delete subset series
                d _2_*

     'Determine the number of f-stats in evaluation
                !no_realruns=0
                for !x=1 to !numberruns
                if fstats(!x+1,1)<>"singular" then
                        !no_realruns=!no_realruns+1
                endif
                next !x

     'Create matrix to calculate summary statistics
                matrix(!no_realruns,2+(2+!otherregs)*2) f_results
                d f_results
                matrix(!no_realruns,2+(2+!otherregs)*2) f_results

                !no_realruns=0
        for !x=1 to !numberruns
                  if fstats(!x+1,1)<>"singular" then

                        !no_realruns=!no_realruns+1
                        f_results(!no_realruns,1)=@val(fstats(!x+1,1))

                         'Determine if f-stat significant or not
                        if @val(fstats(!x+1,2))>0.05 then
                         !isstable=1
                        else
                         !isstable=0
                        endif
                        f_results(!no_realruns,2)=!isstable

                for !y=1 to 2+!otherregs
                                f_results(!no_realruns,2+!y)=@val(fstats(!x+1,9+!y)) 'Store coefficients
```

```
                                        'Determine whether t-stats indicates pos.(1), neg.(-1) or zero (0) coefficient
                                        if @val(fstats(!x+1,10+2+!otherregs+!y))>1.96 then
                                                f_results(!no_realruns,2+2+!otherregs+!y)=1
                                        else
                                                if @val(fstats(!x+1,10+2+!otherregs+!y))<-1.96 then
                                                   f_results(!no_realruns,2+2+!otherregs+!y)=-1
                                                else
                                                   f_results(!no_realruns,2+2+!otherregs+!y)=0
                                                endif
                                        endif
                                next !y
                        endif
        next !x

                'Save values to final output table
                mc_results(!crossindex+1,1)=%current1
                mc_results(!crossindex+1,2)=%current2
                mc_results(!crossindex+1,3)=@mean(@columnextract(f_results,1)) 'mean of f-stats)
                mc_results(!crossindex+1,4)=@mean(@columnextract(f_results,2)) 'share of insignificant f-stats

        for !y=1 to 2+!otherregs

                        mc_results(!crossindex+1,!y+4)=@mean(@columnextract(f_results,!y+2))/@stdev(@columnextra
                        ct(f_results,!y+2))  'mean coef / stdev coef over all runs
                        mc_results(!crossindex+1,!y+4+!otherregs+2)=@mean(@columnextract(f_results,!y+2+!otherregs
                        +2)) 'share of significant coef over all runs

        next !y
                smpl @all

        next !var2

next !var1

{%0}.ls
show mc_results


'Stefan Zeugner 2004
```

Procedure 2: Montecarlo.prg (Eviews)

```
'For %4 times this procedure divides the sample into two distinct subsets by calling draw_rnd_create_eviews3.prg, and fits two separate equations like %0, then performs Chow forecast test between two fits, and saves f-stat, coefs and t-stat into table FSTATS

'EQUATION %0 MUST NOT INCLUDE OPERATORS:
'i.e. no LOG(ser01) or ser01^2 etc.
'Do not use "C" as the constant, use a series containing value 1 instead (e.g. ONESER)

'Objects needed:
'  labels (table): contains at least %2 IDs denoting nodes in 1st column;
'  %0 (equation);
'  %3 (series): weighting series, if no WLS required, give oneser;

'Variables needed:
'%1: nodes per subset; %2: nodes in total sample; %4: number of runs;

'______________________________
'Setting default values for optional parameters
if %0="" then
%0="versuch1"                   'equation to be montecarloed
endif

if %1="" then
  %1="34"                       'number of nodes per subset
endif

if %2="" then                   'total number of nodes
  %2="68"
```

```
endif

if %3="" then
 %3="impdist"                  'weighting series
endif

if %4="" then
  !noruns=10                   'number of montecarlo runs
else
  !noruns=@val(%4)
endif

'Set variables needed for procedure
 !m=@val(%1)
 !n=@val(%2)

 %suffix="_cm"                 'Combined sample denoter
 %subs1="_1_" + %0             'Subsample 1 denoter
 %subs2="_2_" + %0             'Subsample 2 denoter
 !smpllength=!m*(!m+1)/2
 !totallength=!smpllength*2
 !ncoef={%0}.@ncoef

 !smpllength_1=!smpllength+1
 {%0}.makeregs tpregs
 tpregs.add %3
 %regressors="tpregs"

'Create output table
 table fstats
 d fstats
 table fstats

'Label column headers
fstats(1,1)="fstat weighted"
fstats(1,3)="fstat unweighted"
fstats(1,2)="P-Val fstat weighted"
fstats(1,4)="P-Val fstat unweighted"
fstats(1,6)="Adj.R-squared"
fstats(1,7)="subset 1 nodes"

for !i =1 to !ncoef
    fstats(1, 9+!i) = tpregs.@seriesname(!i+1)
next !i




for !run=1 to !noruns

  include draw_rnd_create_eviews3.prg   'routine to create two distinct subsets

  %statusstate =  @str(@round(!run/!noruns*10000)/100) + " Percent"
  statusline %statusstate %status

  {%0}.makeregs temp_regs

    'Create subset 1
    for !i=1 to !ncoef
                    group temp_1reg
                    %nami = temp_regs.@seriesname(!i+1)
                    %name2="_1_"+%nami
                    temp_1reg.add %name2
    next !i

  smpl 1 !smpllength
  stom(temp_1reg,temp_m1reg)

 'Create subset 2
    for !i=1 to !ncoef
                    group temp_2reg
                    %nami = temp_regs.@seriesname(!i+1)
                    %name2="_2_"+%nami
                    temp_2reg.add %name2
    next !i
```

```
  stom(temp_2reg,temp_m2reg)

 if @rank(temp_m1reg) < temp_1reg.@count or @rank(temp_m2reg) < temp_2reg.@count then
   'if subset design matrix is not full rank then goto next run
          fstats(!run+1,1)="singular"
 else

    smpl @all
    'Call subset OLS/WLS procedure
    include meta_subset.prg %0 %1 %2

    'Prepare residual series for Chow test calculation
    smpl 1 !smpllength
    combi.ls(w=cross_w%suffix)
    combi.makeresids temp_1
    combi.ls
    combi.makeresids temp_3

    smpl 1 !totallength
    combi.ls(w=cross_w%suffix)
    combi.makeresids temp_2
    combi.ls
    combi.makeresids temp_4

    combi.ls(w=cross_w%suffix)



    %fstreg= "combi"
    fstats(!run+1,6)=combi.@rbar2
    %selregs=""

    for !i=1 to !m
      !subs1number=vselection(!i)
      %selregs=%selregs + labels(!subs1number) + " "
    next !i
    fstats(!run+1,7)=%selregs


    for !i =1 to !ncoef
     fstats(!run+1, 9+!i) = {%fstreg}.@coefs(!i)
     fstats(!run+1, 10+!ncoef+!i) = {%fstreg}.@tstats(!i)
    next !i




    !i=0
    'F-stats value WLS
    fstats(!run+1,1)=((@sumsq(temp_2)-@sumsq(temp_1))/!smpllength)/(@sumsq(temp_1)/(!smpllength-!ncoef))

    'F-stats prob value WLS
    fstats(!run+1,2)=@fdist(((@sumsq((temp_2))-@sumsq((temp_1)))/!smpllength)/(@sumsq((temp_1))/(!smpllength-
!ncoef)),!smpllength-!ncoef,!smpllength)

    'F-stats value OLS
    fstats(!run+1,3)=((@sumsq(temp_4)-@sumsq(temp_3))/!smpllength)/(@sumsq(temp_3)/(!smpllength-!ncoef))

    'F-stats value OLS
    fstats(!run+1,4)=@fdist(((@sumsq((temp_4))-@sumsq((temp_3)))/!smpllength)/(@sumsq((temp_3))/(!smpllength-
!ncoef)),!smpllength-!ncoef,!smpllength)

    endif
    smpl @all

next !run

d temp_*
'show fstats

'Stefan Zeugner 2004
```

Procedure 3: Draw_Rnd_Create_Eviews3.prg (EViews)

```
'This program draws two different sub-samples for !m bodes out of the total with !n nodes
```

```
'all objects in the first subset are prefixed by "_1_", and "_2_" for the second


'if !m or !n are empty then assign %1 and %2 or define default values
'_______________________________
 series mytest_!m
 series mytest_!n
 group mytest_group mytest_!m mytest_!n
if mytest_group.@seriesname(1)="mytest_" or mytest_group.@seriesname(2)="mytest_" then
  if %1<>"" then 'get routine parameters
    !m=@val(%1)
    !n=@val(%2)
  else    'define default values
    !m=68
    !n=34
  endif
endif
 d mytest_*

if %regressors="" then
   %regressors="tpregs"
endif
'_______________________________

d subset_*
d *_subset

'calculate observation number
!obs=!n*(!n+1)/2

'construct index matrix
matrix(!n,!n) mindex
for !c=1 to !n
		for !r=!c to !n
				mindex(!r,!c)=(2*!n+1-!c)/2*!c-!n+!r
		next !r
next !c


'construct ordered series
series obs=1
smpl 2 @last
obs = obs(-1)+1
smpl @all
series backsort=obs


'DRAW FIRST SUBSET
'draw randomly from series obs
series selection=na
scalar random

for !i=1 to !m
		rndint(random,!n-!i)
		!random=random+1
		selection(!i) = obs(!random)
		for !j=!random to !n
				obs(!j)=obs(!j+1)
		next !j
		%rest=@otod(!n-!i+1)
		obs.fill(o=%rest,l) na
next !i



'write selected values into vector vselection
smpl @all
sort selection
vector(!m) vselection
for !i=1 to !m
		vselection(!i)=selection(!obs-!m+!i)
next !i
sort backsort

'get cross-section indices from index matrix according to vselection
series drawn=na
for !j=1 to !m
		for !i=!j to !m
				!c=vselection(!j)
```

```
			!r=vselection(!i)
			!drawn=mindex(!r,!c)
			drawn(!drawn)=!drawn
		next !i
next !j

series subset_smpl
!id=1
for !i=1 to !obs
		if drawn(!i)<>na then
			subset_smpl(!id)=drawn(!i)
			!id=!id +1
		endif
next !i

'write selected region labels to table subset_labels
table subset_labels
d subset_labels
table subset_labels
subset_labels(1,2)="Subset 1"
subset_labels(1,4)="Subset 2"
for !i=1 to !m
		subset_labels(!i+1,2)=labels(vselection(!i))
next !i

'construct subset group with stacked cross-section data according to vselection

group _1_subset
!noseries={%regressors}.@count
for !i=1 to !noseries
		%temp={%regressors}.@seriesname(!i)
		%subs1="_1_"+%temp

		series temp1=%temp
		series temp2=na
		for !j=1 to !m*(!m+1)/2
			!subset=subset_smpl(!j)
			temp2(!j)=temp1(!subset)
		next !j
		series %subs1
		{%subs1}=temp2
		_1_subset.add {%subs1}
next !i

'DRAW SECOND SUBSET
'draw randomly from series obs
series selection=na
scalar random
if !m>!n/2 then
		statusline Stopped, subset to large!
		stop
Else
		if !m=!n/2 then
		selection=obs
		Else
		for !i=1 to !m
			rndint(random,!n-!m-!i)
			!random=random+1
			selection(!i) = obs(!random)
			for !j=!random to !n
				obs(!j)=obs(!j+1)
			next !j
			%rest=@otod(!n-!m-!i+1)
			obs.fill(o=%rest,l) na
		next !i

		Endif
Endif


'write selected values into vector vselection2
smpl @all
sort selection
vector(!m) vselection2
for !i=1 to !m
		vselection2(!i)=selection(!obs-!m+!i)
next !i
sort backsort
```

```
'get cross-section indices from index matrix according to vselection2
series drawn=na
for !j=1 to !m
	for !i=!j to !m
		!c=vselection2(!j)
		!r=vselection2(!i)
		!drawn=mindex(!r,!c)
		drawn(!drawn)=!drawn
	next !i
next !j

series subset_smpl
!id=1
for !i=1 to !obs
	if drawn(!i)<>na then
		subset_smpl(!id)=drawn(!i)
		!id=!id +1
	endif
next !i

'write selected region labels to table subset_labels
for !i=1 to !m
	subset_labels(!i+1,4)=labels(vselection2(!i))
next !i

'construct subset group with stacked cross-section data according to vselection2

group _2_subset
!noseries={%regressors}.@count
for !i=1 to !noseries
	%temp={%regressors}.@seriesname(!i)
	%subs1="_2_"+%temp
	series temp1=%temp
	series temp2=na
	for !j=1 to !m*(!m+1)/2
		!subset=subset_smpl(!j)
		temp2(!j)=temp1(!subset)
	next !j
	series %subs1
	{%subs1}=temp2
	_2_subset.add {%subs1}
next !i

'set sample to drawn number
%end=@str((!m*(!m+1)/2))
smpl 1 %end

'throw out crap
d obs
d temp*
'd backsort
'show subset_labels
smpl @all
'save subset1.wf1

'Stefan Zeugner 2004
```

Procedure 4: Stacked_Weighting.prg

```
'This procedure performs a WLS with variance weighting based on the sum of observations for specific node (preferably centered observations)
'Provide either a sym object for the weights(%1), or a (stacked) series (%3)

'Set default values for optional parameters

  if %0="" then
	%0="versuch1"		'equation object
  endif
  if %1="" then
	%wlsmatrix="m_impdist"		'sym object for weighting
  else
	%wlsmatrix=%1
  endif
  if %2="" then
	!n=68				'number of nodes
```

```
    else
            !n=@val(%2)
    endif

    if %3="" then
    else
            %current=%3     'construct weighting matrix out of series
            include reversematrix
            %wlsmatrix="mreverse"
    endif

'___________________________
    !obs=!n*(!n+1)/2

    smpl @all
    {%0}.ls
    {%0}.makeresids e

    rowvector(!n) wls_one=1
    vector(!n) wls_vtotcoop=wls_one*{%wlsmatrix}
    mtos(wls_vtotcoop,wls_stotcoop)

    vector(!n) wls_diago


    for !i=1 to !n
                    wls_diago(!i)=wls_vtotcoop(!i)^2
    next !i

    sym wlssigma=@makediagonal(wls_diago)


    for !i=1 to !n
      for !j=!i to !n
            if !i<>!j then
                    wlssigma(!j,!i)=wlssigma(!i,!i)^0.5*wlssigma(!j,!j)^0.5
            endif
      next !j
    next !i

    vector wls_vweights=@vech(wlssigma)
    mtos(wls_vweights,wlsweights)

    d wls_*
    {%0}.ls(w=wlsweights^0.5)
    {%0}.white
```

Procedure 5: Meta_Subset.prg (EViews)

```
'This program is a linker for doing WLS for two subsets _1_ and _2_ and then doing WLS for combined subsets (named "combi")

'Setting default values for optional parameters
if %0="" then
  %0="versuch1"
endif

if %1="" then
  %1="34"
endif

if %2="" then
  %2="68"
endif

for !subscript=1 to 2
  include weightols_subsets_eviews3.prg %0 %1 %2
next !subscript

include combine_chow_eviews3.prg %0 %1 %2

%suffix="_cm"

'Stefan Zeugner 2004
```

Procedure 6: WeightOLS_Subsets_Eviews3.prg (EViews)

```
'This program does a WLS for the subset _%subscipt_

'Set default values for optional parameters
 if %0="" then
        %0="versuch1"
 endif
 if %1="" then
        %1="34"
 endif

 if %weighting="" then
   %weighting=%subscript+"l_geodist"
 endif


 '%subscript="_1_" this line should be activated if this sub-program is run single
 %subscript="_"+@str(!subscript)+"_" ' this line is activated if this sub-program is called by meta_subset.prg

 !m=@val(%1)

 !obs=!m*(!m+1)/2

'Create subset equation object
 smpl 1 !obs
 %subs1=%subscript+%0
 equation %subs1
 d %subs1

'Create subset regressor group
 copy %0 tempreg
 %regs =%subscript+"reg"
 group %regs
 d %regs
 tempreg.makeregs regs
 group %regs
 !noregs=regs.@count

 'Dependent variable
        %nami = regs.@seriesname(1)
        %dep =%subscript+%nami
        %ldep =%subscript+%nami

if %3="" then
        %3=%subscript+%nami
else
        %weighting=%subscript+%3
endif

for !i=2 to !noregs
                %nami = regs.@seriesname(!i)
                %name2=%subscript+%nami

        {%regs}.add %name2
next !i

'DO WLS
sym(!m,!m) mcoop

series weighting=%weighting
for !c=1 to !m
        for !r = !c to !m
                !index=(!m+0.5-!c/2)*!c-!m+!r
                mcoop(!r,!c)=weighting(!index)
        next !r
next !c

d weighting


tempreg.ls %ldep %regs
%e=%subscript+"e"
tempreg.makeresids %e


rowvector(!m) one=1
```

```
vector(!m) totcoop=one*mcoop

vector(!m) diago



!id=1
for !i=1 to !m
		diago(!i)=log(totcoop(!i))^2
		'diago(!i)={%e}(!id)^2
	!id=!m-!i+!id+1
next !i
d totcoop
d one

sym cross_set=@makediagonal(diago)
for !i=1 to !m
	for !j=!i to !m
		if !i<>!j then
			cross_set(!j,!i)=cross_set(!i,!i)^0.5*cross_set(!j,!j)^0.5
		endif
	next !j
next !i

vector vcrossweights=@vech(cross_set)
%cross_weights=%subscript+"cross_weights"
'%hhh=%subscript+"ll_impdist"
series %cross_weights '={%hhh}
mtos(vcrossweights,%cross_weights)

tempreg.ls(w=(%cross_weights))
%u=%subscript+"u"
tempreg.makeresids %u



smpl @all
rename tempreg %subs1
'show %subs1
'Stefan Zeugner 2004
```

Procedure 7: Combine_Chow_Eviews3.prg (EViews)

```
'This program combines the two equations resulting from weightols_subsets_eviews3.prg
'EQUATION %0 MUST NOT INCLUDE OPERATORS
'Objects needed:
'  %0 (equation);


'Setting default values for optional parameters
if %0="" then   'Original parent of the equations to be combined
	%0="versuch1"
endif

if %1="" then  'Number of nodes per subset
	!m=34
else
	!m=@val(%1)
endif

if %2="" then  'Number of nodes in total
	!n=68
else
	!n=@val(%2)
endif

%subs1="_1_"+%0  'name of the first subset equation object

!smpllength=!m*(!m+1)/2
!totallength=!smpllength*2

'lay out series and variables needed
group reg_combined
%suffix="_cm"
```

```
d *%suffix
d reg_combined
group reg_combined


!seriesno=_1_reg.@count 'number of regressors
for !i=1 to !seriesno

  'Create combined series
                    %temp1=_1_reg.@seriesname(!i)  'name of series !i
                    %temp2=@mid(%temp1,4,40)+%suffix
                    series temp1=%temp1

                    smpl 1 !totallength
                    %temp3= _2_reg.@seriesname(!i)
                    series temp2=%temp3
                    for !j=!smpllength+1 to !totallength
                            temp1(!j)=temp2(!j-!smpllength)
                    next !j

                    copy temp1 %temp2
                    reg_combined.add %temp2 'add combined series with suffix "_cm" and add to reg_combined

next !i

    'Stack series created out of subset 1 with obs of subset 2
            {%subs1}.makeregs temp4
            %temp5= temp4.@seriesname(1)
                    %temp7="_2_"+@mid(%temp5,4,40)
                    %temp5=%temp5
                    %temp6=@mid(%temp5,4,40) + %suffix

            series temp7=%temp7
            series temp6=%temp5
            for !j=!smpllength+1 to !totallength
                            temp6(!j)=temp7(!j-!smpllength)
            next !j

            rename temp6 %temp6


            series cross_w%suffix=_1_cross_weights
            for !j=!smpllength+1 to !totallength
                            cross_w%suffix(!j)=_2_cross_weights(!j-!smpllength)
            next !j

'Do combined WLS
smpl 1 !totallength
equation combi.ls(w=1/cross_w%suffix) %temp6 reg_combined
combi.makeresids u%suffix
!smpllength_1=!smpllength+1
'combi.chow(f) !smpllength_1
d temp*
smpl @all

'Stefan Zeugner 2004
```

Multi-dimensional Scaling Procedure for Distance Matrix

This is the procedure used for the creation of the “maps” in Figure 4, Figure 5 and Figure 6. Its analytic base is described in section 5.1.2, p. 88. By multi-dimensional scaling, it transforms a symmetric matrix of distances into a matrix of coordinates. The resulting coordinate vectors are ranked by importance from left to right. The parameters or the default values in the code define the dimensions the input matrix and the path of the excel file containing the input matrix and receiving the output matrix.

Procedure 8: Scale.prg

```
'This file provides multi-dimensional scaling for the coordinates derived from a distance matrix saved under %excelfilepath (see section 5.1.2 in the diploma thesis). Resulting dimensions are in columns and ordered by importance (i.e. abs(eigenvalue))

'__ Optional Parameters: ___________________
  '%0: number of nodes
  '%1: number of dimensions
  '%2: Path of the excel file with distance matrix
'___________________________________________


'__ Setting default values for parameters ________
  !n=68
  !dim=68
  %excelfilepath="c:\data\distance.xls"
'___________________________________________


  if %0="" then
    !n=@val(%0)
  endif
  if %1="" then
    !dim=@val(%1)
  endif
  if %2="" then
    %excelfilepath=%2
  endif

'___________________________________________


  if !dim>!n then
     stop
  endif

  'load dist matrix
  create distance u 1 !n
  sym(!n) dist_mat
  dist_mat.read(b2, t=xls) %excelfilepath !n

  'transform distance data
  sym(!n) A

  for !i=1 to !n
     for !j=1 to !n
        A(!i,!j)=-0.5*dist_mat(!i,!j)^2
     next !j
  next !i

  vector(!n) one=1
  sym H=@identity(!n)-one*@transpose(one)/!n
  sym B=H*A*H

  'Get eigenvector decomposition
  matrix eivec=@eigenvectors(B)
  vector eival=@eigenvalues(B)

  'Sort by most important eigenvalues
  series backsort=na
  for !i=1 to !n
     backsort(!i)=!i
  next !i
  mtos(eival,eivals)
  group coords
  sort @abs(eivals)

  for !l=1 to !dim
        !index=backsort(!n-!i+1)
        if eivals(!n-!i+1) >0 then
                              !eisign=1
        else
           !eisign=-1
        endif

       vector vcoord_!i=!eisign*@abs(eivals(!n-!i+1))^0.5*@columnextract(eivec,!index)
        mtos(vcoord_!i,coord_!i)
        coords.add coord_!i
  next !i

'do output and write it into %excelfilepath
```

```
  write(t=xls) %excelfilepath coords
  show coords
  stop
  close distance

'Stefan Zeugner 2004
```

Eviews procedures for matrix transformation

The following procedures facilitate transformation between matrices and stacked series. Procedure 9 takes a series stacked from the observations in the main diagonal and lower triangular of a symmetric matrix (which can be constructed with the EViews command @vech). It then re-constructs the original matrix to which the @vech command was applied.

Procedure 10 stacks a lower triangular matrix without taking into account the main diagonal (as opposed to @vech).For instance this was needed in compiling the 2278 observation series for Estimation 2 and Estimation 3. Procedure 11 re-transforms such a series into a symmetric matrix with an exogenously provided main diagonal.

Procedure 9: Reversematrix.prg (EViews)

```
'This procedure converts a stacked series of n*(n+1)/2 observations into a symmetric n x n matrix

'set default values for optional parameters
 if %0="" then
   %0="coopfp4"
 endif
 if %2="" then
   !n=68
 else
   !n=@val(%2)
 endif

 if %current="" then
   %current=%0
 endif
 sym(!n,!n) mreverse

for !c=1 to !n
		for !r=!c to !n
				!index=(2*!n+1-!c)/2*!c-!n+!r
				mreverse(!r,!c)={%current}(!index)
		next !r
next !c
'show mreverse
```

Procedure 10: Lower_Vech.prg (EViews)

```
'Stacks the lower triangular of a symmetric matrix into series (i.e. without the main diagonal)

'%0: series name
'%1: matrix name

 if %1="" then
   %1 ="m_" + %0
 endif

 !n=@columns({%1})
 !nbobs=!n*(!n-1)/2

 %vect = "v_" + %0
 vector(!nbobs) {%vect}

 !index=1
for !j=1 to !n
 for !i=!j+1 to !n
```

```
    {%vect}(!index) = {%1}(!i,!j)
    !index=!index+1
  next !i
next !j

  mtos(%vect,%0)
```

Procedure 11: Restack_lower_vech.prg (EViews)

```
'This procedure transforms a (lower-triangular) stacked series back to a symmetric matrix. I.e. a uniform value for the main diagonal has to be provided (%2, default=1)

Paramters & Object needed:
'%0: stacked series object
'%1: n, i.e. number of rows/columns of the nxn output matrix
'%2: uniform value for the main diagonal

if %2 ="" then
  !diago=1
else
  !diago=@val(%2)
endif

  smpl @all
  %ser=%0
  !n=@val(%1)

  if @obs({%ser})<>(!n^2-!n)/2 then
    'if series does not match given n then stop
    statusline Wrong n
    stop
  endif
  !index=1
  %mat="rm_" + %ser
  sym(!n) {%mat}

'fill matrix
for !j=1 to !n
  for !i=!j+1 to !n
    {%mat}(!i,!j)={%ser}(!index)
    !index=!index+1
  next !i
next !j

'fill main diagonal
for !i=1 to !n
  {%mat}(!i,!i)=!diago
next !i
```

VBA procedure for scaling estimated interaction matrix by row sums

The following two procedures interact in order to scale a collaboration matrix as in section 6.1.5: It starts by calling the Sub *Show_doubleconstrained_form()* which displays *UserForm1*. Pressing the *Run Iterations* button performs the scaling with input provided in the boxes *borders*, *input matrix* and *precision* – by calling first Sub *IterButton_Click()* and then Sub *Double_Constr_Conditioning()*. Finally pressing *PasteButton_Click()* inserts the values into the area defined in *output matrix*.

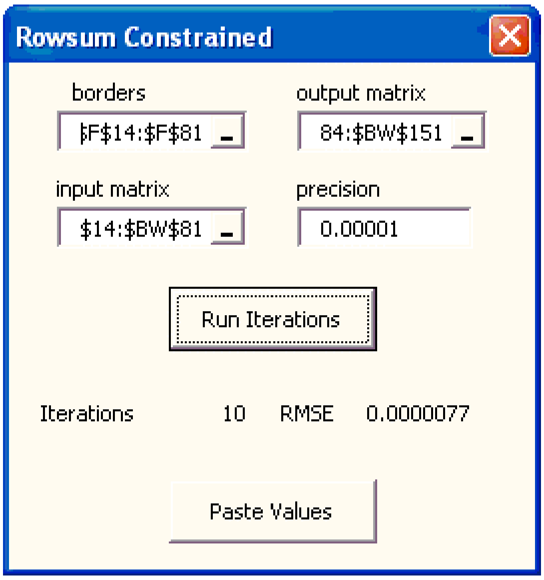
Procedure 12: Dipla_VBA.xls!Module Dipla (Excel VBA)

```
Public CoopArray
```

```
Public Sub Show_doubleconstrained_form()
  'Display UserForm1
  With UserForm1
    .ValRMSE.Visible = False
    .LabelRMSE.Visible = False
    .ValIter.Visible = False
    .LabelIter.Visible = False
    .PasteButton.Visible = False
    .TextBox1.Value = "0.00001"
    .Show
  End With
End Sub

Sub Double_Constr_Conditioning(ByVal borderrange, ByVal cooprange, ByVal output, Precision As Double)
   'This procedure is called by UserForm1.IterButton_Click()

   Dim Border, SumVector, HorSum, ScaleMatrix, IniArray, FinalScaleMatrix
   Dim oneVector() As Double
   Dim sqerror As Double
   Dim i As Long, j As Long, l As Long, n As Long



   'Transfer ranges to arrays
   Border = borderrange
   CoopArray = cooprange
   IniArray = CoopArray

   If Not UBound(Border, 1) = UBound(CoopArray, 1) Then
      MsgBox "n not identifiable", vbExclamation
      Exit Sub
   ElseIf Not UBound(CoopArray, 1) = UBound(CoopArray, 2) Then
      MsgBox "n not identifiable", vbExclamation
      Exit Sub
   End If

   n = UBound(Border, 1)
   'Create 1-Vector
   ReDim oneVector(1 To n, 1 To 1)
   For i = 1 To n
      oneVector(i, 1) = 1
   Next i

   'Create unity matrix and a diagonal matrix
   ReDim ScaleMatrix(1 To n, 1 To n)
   ReDim FinalScaleMatrix(1 To n, 1 To n)
   For j = 1 To n
      For i = 1 To n
         If i = j Then FinalScaleMatrix(i, j) = 1 Else FinalScaleMatrix(i, j) = 0
         ScaleMatrix(i, j) = 0
      Next i
   Next j

   '=============================
   'Iterate to Convergence (defined by Precision)

   l = 1
   Do

      SumVector = Application.MMult(Application.Transpose(oneVector), CoopArray)

      For i = 1 To n
         ScaleMatrix(i, i) = Border(i, 1) / SumVector(i)
      Next i

      CoopArray = Application.MMult(ScaleMatrix, Application.Transpose(CoopArray))
      FinalScaleMatrix = Application.MMult(ScaleMatrix, FinalScaleMatrix)


      l = l + 1
      sqerror = 0
      For i = 1 To n
         sqerror = sqerror + (SumVector(i) - Border(i, 1)) ^ 2
      Next i

   Loop Until (sqerror / n) ^ 0.5 < Precision

   With UserForm1
      .ValIter.Caption = l
```

```
        .ValRMSE.Caption = Int((sqerror / n) ^ 0.5 * 10000000) / 10000000
    End With

   'Continue with UserForm1
End Sub
```

Procedure 13: Dipla_VBA.xls!UserForm UserForm1 (Excel VBA)




```
Private Sub IterButton_Click()
    'This procedure is called by pressing "Run Iterations"
    Call double_constr_conditioning(Range(RefEdit1.Value), Range(RefEdit2.Value), RefEdit3.Value, Val(TextBox1.Value))

    With UserForm1
        .ValRMSE.Visible = True
        .LabelRMSE.Visible = True
        .ValIter.Visible = True
        .LabelIter.Visible = True
        .PasteButton.Visible = True
    End With
End Sub

Private Sub PasteButton_Click()
    'This procedure is called by pressing "Paste Values"
    Range(RefEdit3.Value) = CoopArray
    Range(RefEdit3.Value).Select
    UserForm1.Hide

End Sub
```