Regional Innovation Scoreboard 2012

Methodology report

UNU-MERIT - Hugo Hollanders
Cambridge Econometrics - James Derbyshire, Richard Lewney
CWTS - Robert Tijssen
Joint Research Centre - Stefano Tarantola
Technopolis - Lorena Rivera Leon

June 2012

This report accompanies the
Regional Innovation Scoreboard (RIS) 2012 report

Table of Contents

1. Introduction 3
2. Regional innovation indicators 4
2.1 Innovation Union Scoreboard: regional data availability 4
2.2 Indicator definitions 6
2.3 Community Innovation Survey data 14
3. Methodology 16
3.1 Imputation of missing data 16
3.2 Normalisation of the indicators 16
4. Robustness test of clustering methods 18
4.1 Clustering methods 19
4.1.1 Hierarchical cluster analysis 19
4.1.2 Squared Euclidean Distance 20
4.1.3 Non-hierarchical (k-means) cluster analysis 20
4.2 Robustness tests 20
4.2.1 Hierarchical cluster analysis using Ward's method 21
4.2.2 Hierarchical cluster analysis using Average 'between-groups' linkage and Squared Euclidean Distance 24
4.2.3 Non-hierarchical k-means cluster analysis 25
4.3 Conclusions cluster analysis 26
5. Conclusions for RIS 2012 28
Annex A: Allocation of cases to clusters under different clustering methods 31
Annex B: Calculation of Regional Innovation Index 35

1. Introduction

Innovation is a key factor determining productivity growth. Understanding the sources and patterns of innovative activity in the economy is fundamental to develop better policies. The Innovation Union Scoreboard (IUS) benchmarks on a yearly basis the innovation performance of Member States, drawing on statistics from a variety of sources, including the Community Innovation Survey. It is increasingly used as a reference point by innovation policy makers across the EU.

The IUS benchmarks performance at the level of Member States, but innovation plays an increasing role in regional development, both in the Lisbon strategy and in Cohesion Policy. Regions are important engines of economic development. Geographical proximity matters in business performance and in the creation of innovation. Recognising this, innovation policy is increasingly designed and implemented at regional level. However, despite some advances, there is an absence of regional data on innovation indicators which could help regional policy makers design and monitor innovation policies.

The Regional Innovation Scoreboard (RIS) addresses this gap and provides statistical facts on regions' innovation performance. Following the revision of the IUS in 2010, the RIS 2012 will use as many of the IUS indicators at the regional level for all EU Member States any other countries participating in the EIP for which sufficient data is available.

This methodological report examines the available data and discusses how they can be used to develop a Regional Innovation Scoreboard. Section 2 will discuss regional data availability of the IUS indicators and will provide detailed definitions for each of the indicators. Section 3 will explain the imputation technique for missing data and will discuss the methodology for normalising the data. Section 4 discusses different cluster methods which can be used for determining a typology of regions. Section 5 concludes by summarizing the indicators that will be used in the RIS 2012.

2. Regional innovation indicators

2.1 Innovation Union Scoreboard: regional data availability

The Innovation Union Scoreboard (IUS) uses a wide variety of indicators to measure innovation performance at the country level and distinguishes between 3 main types of indicators and 8 innovation dimensions, capturing in total 25 different indicators. The indicators are grouped into dimensions to capture different aspects of innovation performance as summarized in Table 1. The IUS 2010 Methodology Report explains the rationale for the selection of indicators and for their grouping[1].

Table 1: Innovation dimensions used in the Innovation Union Scoreboard

- The **Enablers** capture the main drivers of innovation performance external to the firm and it differentiates between 3 innovation dimensions:
  - **"Human resources"** includes 3 indicators and measures the availability of a high-skilled and educated workforce.
  - **"Open, excellent and attractive research systems"** includes 3 indicators and measures the international competitiveness of the science base.
  - **"Finance and support"** includes 2 indicators and measures the availability of finance for innovation projects and the support of governments for research and innovation activities.
- **Firm activities** capture the innovation efforts at the level of the firm and it differentiates between 3 innovation dimensions:
  - **"Firm investments"** includes 2 indicators of both R&D and non-R&D investments that firms make in order to generate innovations.
  - **"Linkages & entrepreneurship"** includes 3 indicators and measures entrepreneurial efforts and collaboration efforts among innovating firms and also with the public sector.
  - **"Intellectual assets"** captures different forms of Intellectual Property Rights (IPR) generated as a throughput in the innovation process.
- **Outputs** capture the effects of firms' innovation activities and it differentiates between 2 innovation dimensions.
  - **"Innovators"** includes 3 indicators and measures the number of firms that have introduced innovations onto the market or within their organisations, covering both technological and non-technological innovations and the presence of high-growth firms.
  - **"Economic effects"** includes 5 indicators and captures the economic success of innovation in employment, exports and sales due to innovation activities.

Table 2 summarizes data availability at the regional level for the indicators used in IUS. Of the 24 indicators used in IUS, data at the regional level are only available for 12 indicators. Data availability differs between the different innovation dimensions. For Enablers data availability is poor with only 1 indicator for Human resources and Finance and support and no indicator for Open, excellent and attractive research systems. For Firm activities data availability is good, in particular for Firm investments and Linkages & entrepreneurship with data for all indicators. For Intellectual assets regional data are available for 1 out of 4 indicators. For Outputs data availability is very good for Innovators and good for Economic effects.

[1] Hollanders, H. and S. Tarantola (2010), "Innovation Union Scoreboard 2010 - Methodology report", INNO Metrics Thematic Paper, Brussels: DG Enterprise and Industry.

Table 2: Innovation Union Scoreboard: regional data availability

IUS indicators	Regional data availability
Human resources	
1.1.1 New doctorate graduates (ISCED 6) per 1000 population aged 25-34	No
1.1.2 Percentage population aged 30-34 having completed tertiary education	Percentage population aged 25-64 having completed tertiary education
1.1.3 Percentage youth aged 20-24 having attained at least upper secondary level education	No
Open, excellent and attractive research systems	
1.2.1 International scientific co-publications per million population	No
1.2.2 Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country	No
1.2.3 Non-EU doctorate students as a % of all doctorate students	No
Finance and support	
1.3.1 R&D expenditure in the public sector as % of GDP	Identical
1.3.2 Venture capital (early stage, expansion and replacement) as % of GDP	No
Firm investments	
2.1.1 R&D expenditure in the business sector as % of GDP	Identical
2.1.2 Non-R&D innovation expenditures as % of turnover	Similar (only for SMEs)
Linkages & entrepreneurship	
2.2.1 SMEs innovating in-house as % of SMEs	Identical
2.2.2 Innovative SMEs collaborating with others as % of SMEs	Identical
2.2.3 Public-private co-publications per million population	Identical
Intellectual assets	
2.3.1 PCT patent applications per billion GDP (in PPP€)	EPO patent applications per billion GDP (in PPP€)
2.3.2 PCT patent applications in societal challenges per billion GDP (in PPP€) (climate change mitigation; health)	No
2.3.3 Community trademarks per billion GDP (in PPP€)	No
2.3.4 Community designs per billion GDP (in PPP€)	No
Innovators	
3.1.1 SMEs introducing product or process innovations as % of SMEs	Identical
3.1.2 SMEs introducing marketing or organisational innovations as % of SMEs	Identical
3.1.3 High-growth innovative firms	N/A
Economic effects	
3.2.1 Employment in knowledge-intensive activities (manufacturing and services) as % of total employment	Employment in knowledge-intensive services as % of total employment Employment in medium-high and high-tech manufacturing as % of total workforce
3.2.2 Medium and high-tech product exports as % total product exports	No
3.2.3 Knowledge-intensive services exports as % total service exports	No
3.2.4 Sales of new to market and new to firm innovations as % of turnover	Similar (only for SMEs)
3.2.5 License and patent revenues from abroad as % of GDP	No

2.2 Indicator definitions

This section provides detailed definitions for each of the indicators used in RIS. The numbering of the indicators follows that used in IUS (cf. Table 2).

1.1.2 Population with tertiary education per 100 population aged 25-64	
Numerator	Number of persons in age class with some form of post-secondary education (ISCED 5 and 6)
Denominator	The reference population is all age classes between 25 and 64 years inclusive
Rationale	This is a general indicator of the supply of advanced skills. It is not limited to science and technical fields because the adoption of innovations in many areas, in particular in the service sectors, depends on a wide range of skills. Furthermore, it includes the entire working age population, because future economic growth could require drawing on the non-active fraction of the population. International comparisons of educational levels however are difficult due to large discrepancies in educational systems, access, and the level of attainment that is required to receive a tertiary degree. Differences among countries should be interpreted with caution
Included in RIS 2009	Yes
Included in IUS	Comparable, IUS refers to age group 30-34
Data source	Eurostat
Data availability	NUTS 2, 2000-2009

1.3.1 Public R&D expenditures (% of GDP)	
Numerator	All R&D expenditures in the government sector (GOVERD) and the higher education sector (HERD). Both GOVERD and HERD according to the Frascati-manual definitions, in national currency and current prices
Denominator	Gross Domestic Product, in national currency and current prices
Rationale	R&D expenditure represents one of the major drivers of economic growth in a knowledge-based economy. As such, trends in the R&D expenditure indicator provide key indications of the future competitiveness and wealth of the EU. Research and development spending is essential for making the transition to a knowledge-based economy as well as for improving production technologies and stimulating growth
Included in RIS 2009	Yes
Included in IUS	Yes
Data source	Eurostat
Data availability	2000 - ...: NUTS 1: BE (2007), PL (2008) NUTS 2: BG (2008), CZ (2008), DE (2007), IE (2008), GR (2005), ES (2008), FR (2004), IT (2007), HU (2008), NL (2007), AT (2007), PL (2007), PT (2008), RO (2008), SI (2008), SK (2008), FI (2008), SE (2007), UK (2008) NUTS 3: DK (2007)

2.1.1 Business R&D expenditures (% of GDP)	
Numerator	All R&D expenditures in the business sector (BERD), according to the Frascati-manual definitions, in national currency and current prices
Denominator	Gross Domestic Product, in national currency and current prices
Rationale	The indicator captures the formal creation of new knowledge within firms. It is particularly important in the science-based sector (pharmaceuticals, chemicals and some areas of electronics) where most new knowledge is created in or near R&D laboratories
Included in RIS 2009	Yes
Included in IUS	Yes

Data source	Eurostat
Data availability	2000 - ...: NUTS 1: BE (2007), PL (2008) NUTS 2: BG (2008), CZ (2008), DE (2007), IE (2008), GR (2005), ES (2008), FR (2004), IT (2007), HU (2008), NL (2007), AT (2007), PL (2007), PT (2008), RO (2008), SI (2008), SK (2008), FI (2008), SE (2007), UK (2008) NUTS 3: DK (2007)

2.1.2 Non-R&D innovation expenditures (% of total turnover)		
Numerator	Sum of total innovation expenditure for SMEs only, in national currency and current prices excluding intramural and extramural R&D expenditures	
Denominator	Total turnover for SMEs only (both innovators and non-innovators), in national currency and current prices	
Rationale	This indicator measures non-R&D innovation expenditure as percentage of total turnover. Several of the components of innovation expenditure, such as investment in equipment and machinery and the acquisition of patents and licenses, measure the diffusion of new production technology and ideas. Compared to the EIS 2007 the indicator no longer captures intramural and extramural R&D expenditures and thus no longer overlaps with the indicator on business R&D expenditures	
Included in RIS 2009	Yes	
Included in IUS	Yes, but for all firms	
Data source	Community Innovation Survey - Eurostat in collaboration with Member States	
Data availability	AT: NUTS 1 2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2008 NO: NUTS 2 2004-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2004-2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008

2.2.1 SMEs innovating in-house (% of all SMEs)	
Numerator	Sum of SMEs with in-house innovation activities. Innovative firms with in-house innovation activities have introduced a new product or new process either in-house or in combination with other firms. The indicator does not include new products or processes developed by other firms
Denominator	Total number of SMEs (both innovators and non-innovators).
Rationale	This indicator measures the degree to which SMEs, that have introduced any new or significantly improved products or production processes during the period 2002-2004, have innovated in-house. The indicator is limited to SMEs because almost all large firms innovate and because countries with an industrial structure weighted to larger firms would tend to do better
Included in RIS 2009	Yes
Included in IUS	Yes
Data source	Community Innovation Survey - Eurostat in collaboration with Member States

Data availability	AT: NUTS 1 2004-2006-2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FI: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2006-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2004-2008 NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2004-2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008 UK: NUTS 1 2004-2006

2.2.2 Innovative SMEs collaborating with others (% of all SMEs)		
Numerator	Sum of SMEs with innovation co-operation activities. Firms with co-operation activities are those that had any co-operation agreements on innovation activities with other enterprises or institutions in the three years of the survey period	
Denominator	Total number of SMEs	
Rationale	This indicator measures the degree to which SMEs are involved in innovation co-operation. Complex innovations, in particular in ICT, often depend on the ability to draw on diverse sources of information and knowledge, or to collaborate on the development of an innovation. This indicator measures the flow of knowledge between public research institutions and firms and between firms and other firms. The indicator is limited to SMEs because almost all large firms are involved in innovation co-operation	
Included in RIS 2009	Yes	
Included in IUS	Yes	
Data source	Community Innovation Survey - Eurostat in collaboration with Member States	
Data availability	AT: NUTS 1 2004-2006-2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FI: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2006-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2004-2008 NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2004-2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008 UK: NUTS 1 2004-2006

2.2.3 Public-private co-publications	
Numerator	Number of public-private co-authored research publications (PPCs). The definition of the "private sector" covers business enterprises and for-profit organizations, but excludes the private medical and health sector. Publications are assigned to the region in which the private sector organization is physically located.
Denominator	Total population or total publication output
Rationale	This indicator captures public-private research linkages and active collaboration activities between business sector researchers and public sector researchers resulting in academic publications
Included in RIS 2009	No
Included in IUS	Yes
Data source	CWTS (Web of Science database)
Data availability	NUTS 2 (all regions with sufficiently large PPC output), 2007-2008

2.3.1 EPO patent applications per billion GDP (in PPP€)	
Numerator	Number of patents applied for at the European Patent Office (EPO), by year of filing. The national distribution of the patent applications is assigned according to the address of the inventor
Denominator	Gross Domestic Product in Purchasing Power Parity Euros
Rationale	The capacity of firms to develop new products will determine their competitive advantage. One indicator of the rate of new product innovation is the number of patents. This indicator measures the number of patent applications at the European Patent Office
Included in RIS 2009	Yes
Included in IUS	No, IUS uses PCT patent applications (per billion GDP)
Data source	Eurostat
Data availability	NUTS 2: 2000-2007

3.1.1 Technological (product or process) innovators (% of all SMEs)		
Numerator	The number of SMEs who introduced a new product or a new process to one of their markets	
Denominator	Total number of SMEs	
Rationale	Technological innovation as measured by the introduction of new products (goods or services) and processes is key to innovation in manufacturing activities. Higher shares of technological innovators should reflect a higher level of innovation activities	
Included in RIS 2009	Yes	
Included in IUS	Yes	
Data source	Community Innovation Survey - Eurostat in collaboration with Member States	
Data availability	AT: NUTS 1 2004-2006-2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FI: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2006-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2004-2008 NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2004-2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008 UK: NUTS 1 2004-2006

3.1.2 Non-technological (marketing or organisational) innovators (% of all SMEs)	
Numerator	The number of SMEs who introduced a new marketing innovation and/or organisational innovation to one of their markets
Denominator	Total number of SMEs
Rationale	The Community Innovation Survey mainly asks firms about their technical innovation. Many firms, in particular in the services sectors, innovate through other non-technological forms of innovation. Examples of these are organisational innovations. This indicator tries to capture the extent that SMEs innovate through non-technological innovation
Included in RIS 2009	Yes
Included in IUS	Yes
Data source	Community Innovation Survey - Eurostat in collaboration with Member States

Data availability	AT: NUTS 1 2004-2006-2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FI: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2006-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2004-2008 NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2004-2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008 UK: NUTS 1 2004-2006

3.2.1a Employment in knowledge-intensive services (% of total workforce)	
Numerator	Number of employed persons in the knowledge-intensive services sectors. These include water transport (NACE 61), air transport (NACE 62), post and telecommunications (NACE64), financial intermediation (NACE 65), insurance and pension funding (NACE 66), activities auxiliary to financial intermediation (NACE 67), real estate activities (NACE 70), renting of machinery and equipment (NACE 71), computer and related activities (NACE72), research and development (NACE73) and other business activities (NACE 74)
Denominator	Total workforce including all manufacturing and service sectors
Rationale	Knowledge-intensive services provide services directly to consumers, such as telecommunications, and provide inputs to the innovative activities of other firms in all sectors of the economy. The latter can increase productivity throughout the economy and support the diffusion of a range of innovations, in particular those based on ICT
Included in RIS 2009	Yes
Included in IUS	No (IUS uses indicator on employment in knowledge-intensive activities)
Data source	Eurostat
Data availability	NUTS 2: 2000-2008

3.2.1b Employment in medium-high and high-tech manufacturing (% of total workforce)	
Numerator	Number of employed persons in the medium-high and high-tech manufacturing sectors. These include chemicals (NACE24), machinery (NACE29), office equipment (NACE30), electrical equipment (NACE31), telecommunications and related equipment (NACE32), precision instruments (NACE33), automobiles (NACE34) and aerospace and other transport (NACE35)
Denominator	Total workforce including all manufacturing and service sectors
Rationale	The share of employment in high technology manufacturing sectors is an indicator of the manufacturing economy that is based on continual innovation through creative, inventive activity. The use of total employment gives a better indicator than using the share of manufacturing employment alone, since the latter will be affected by the hollowing out of manufacturing in some countries
Included in RIS 2009	Yes
Included in IUS	No (IUS uses indicator on employment in knowledge-intensive activities)
Data source	Eurostat
Data availability	NUTS 2: 2000-2008

3.2.4a Sales of new-to-market products (% of total turnover)	
Numerator	Sum of total turnover of new or significantly improved products for SMEs only
Denominator	Total turnover for SMEs only (both innovators and non-innovators), in national currency and current prices
Rationale	Community Innovation Survey - Eurostat in collaboration with Member States
Included in RIS 2009	Yes

Included in IUS	No, merged with indicator on sales of new-to-firm products	
Data source	Community Innovation Survey Eurostat in collaboration with Member States – CONFIDENTIAL	
Data availability	AT: NUTS 1 2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008

3.2.4b Sales of new-to-firm products (% of total turnover)		
Numerator	Sum of total turnover of new or significantly improved products to the firm but not to the market for SMEs only	
Denominator	Total turnover for SMEs only (both innovators and non-innovators), in national currency and current prices	
Rationale	This indicator measures the turnover of new or significantly improved products to the firm as a percentage of total turnover. These products are not new to the market. Sales of new to the firm but not new to the market products are a proxy of the use or implementation of elsewhere already introduced products (or technologies). This indicator is a proxy for the degree of diffusion of state-of-the-art technologies	
Included in RIS 2009	Yes	
Included in IUS	No, merged with indicator on sales of new-to-firm products	
Data source	Community Innovation Survey - Eurostat in collaboration with Member States	
Data availability	AT: NUTS 1 2008 BE: NUTS 1 2004-2006-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008

In addition to the indicators used in IUS, the following indicators are also included in the robustness test of clustering methods in section 4 as these indicators were also included in the RIS 2009.

Households with broadband access ("Finance and support)	
Numerator	Number of households with broadband access
Denominator	Total number of households
Rationale	Realising Europe's full e-potential depends on creating the conditions for electronic commerce and the Internet to flourish. This indicator captures the relative use of this e-potential by the number of households that have access to broadband
Included in RIS 2009	Yes
Included in IUS	No
Data source	Eurostat
Data availability	NUTS 2: 2006-2010

Resource efficiency innovators (% of all SMEs) ("Innovators") Average of the following 2 indicators:		
	Reduced labour costs resulting from process innovations (% of SMEs)	
Numerator	Sum of innovating SMEs who replied that their product or process innovation had a highly important effect on reducing materials and energy per unit of output	
Denominator	Total number of SMEs	
Rationale	This indicator captures the cost savings from process innovation	
	Reduced use materials and energy resulting from process innovations (% of SMEs)	
Numerator	Sum of innovating SMEs who replied that their product or process innovation had a highly important effect on reducing materials and energy per unit of output	
Denominator	Total number of SMEs	
Rationale	This indicator captures the energy savings from process innovation	
Included in RIS 2009	Yes	
Included in IUS	No	
Data source	Community Innovation Survey - Eurostat in collaboration with Member States	
Data availability	AT: NUTS 1 2004-2006-2008 BE: NUTS 1 2004-2008 BG: NUTS 1 2004-2006-2008 CZ: NUTS 2 2004-2006-2008 ES: NUTS 2 2004-2006-2008 FI: NUTS 2 2004-2006-2008 FR: NUTS 1 2004-2006-2008 GR: NUTS 2 2006 HU: NUTS 2 2006-2008	IT: NUTS 2 2004-2008 NO: NUTS 2 2004-2006-2008 PL: NUTS 2 2004-2006-2008 PT: NUTS 2 2006-2008 RO: NUTS 2 2004-2006-2008 SE: NUTS 2 2008 SI: NUTS 2 2004-2006-2008 SK: NUTS 2 2004-2006-2008 UK: NUTS 1 2004-2006

In addition to the indicators used in IUS, the following indicators are also included in the robustness test of clustering methods in section 4 as these indicators capture a relevant aspect innovation performance.

'Attitude to Entrepreneurship' ("Finance and support")	
Numerator	Average score of: • Important to try new and different things in life from 1 = "very much like me" to 6 = "Not like me at all" • Important to think new ideas and being creative from 1 = "very much like me" to 6 = "Not like me at all"
Rationale	Attitude to new things creates favourable conditions for entrepreneurship
Included in RIS 2009	No
Included in IUS	No
Data source	European Social Survey
Data availability	NUTS 2: 2008

Capital stock data per million population ("Firm investments")	
Numerator	No numerator or denominator, simply an estimate.
Denominator	No numerator or denominator, simply an estimate.
Rationale	New product innovation requires new capital to produce the new product. New processes associated with product innovation require additional or new capital
Included in RIS 2009	No
Included in IUS	No

Data source	Cambridge Econometrics
Data availability	There is an estimate available for every EU NUTS-2 region by three sectors and five assets types. Work has been conducted to disaggregate into six sectors but the data is less robust at this level of disaggregation. An aggregate NUTS-2 level estimate is also available. The data currently run to 2009. For some countries (notably Bulgaria) there was little on which to create a base-year capital stock for 1995 from which to begin the perpetual inventory method whereby investment data is added and deducted (in other words, the data is more robust for some countries than for others). For these countries, such as Bulgaria, it was necessary to produce an initial capital-stock estimate based on that of 'similar' countries.

The following two indicators on structural fund allocations will be used for additional analyses in the RIS report.

Structural Fund allocations on core RTDI activities in the 2007-2013 programming period per million population ("Finance and support")	
Numerator	Core RTDI has been defined using the following fields of intervention (FOIs): 01: R&TD activities in research centres; 02: R&TD infrastructure and centres of competence in a specific technology; 03: Technology transfer and improvement of cooperation networks; 04. Assistance to R&TD, particularly in SMEs (including access to R&TD services in research centres); 74. Developing human potential in the field of research and innovation, in particular through postgraduate studies
Denominator	Total population
Rationale	Reflects the endowment of a region with public financial resources. The SF contribute to activities with a direct link to RTDI
Included in RIS 2009	No
Included in IUS	No
Data source	DG REGIO / Technopolis
Data availability	NUTS 2

Structural Fund allocations on business innovation in the 2007-2013 programming period per million population ("Finance and support")	
Numerator	"Business innovation" has been defined using the following fields of intervention (FOIs): 05: Advanced support services for firms and groups of firms; 06: Assistance to SMEs for the promotion of environmentally-friendly products and production processes; 07: Investment in firms directly linked to research and innovation; 08: Other investment in firms; 09. Other measures to stimulate research and innovation and entrepreneurship in SMEs; 14: Services and applications for SMEs (e-commerce, education and training, networking, etc.)
Denominator	Total population
Rationale	Reflects the endowment of a region with public financial resources. The SF contribute to activities where the link is weaker or indirect e.g. business advisory services with a focus on organisational or technical improvements to the way enterprises operate, that may or may not involve innovation
Included in RIS 2009	No
Included in IUS	No
Data source	DG REGIO / Technopolis
Data availability	NUTS 2

2.3 Community Innovation Survey data

Regional data from the Community Innovation Survey are not directly available from most countries. As for the RIS 2009, these data have been collected directly from the Member States by Eurostat for 2 types of indicators. CIS 2004 and 2006 data have been collected for RIS 2009 and CIS 2008 data have been collected for RIS 2012.

Share-based indicators

These indicators calculate the share of certain innovating firms out of the total population of firms:

- SMEs innovating in-house (% of all SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BE, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

- Innovative SMEs collaborating with others (% of all SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BE, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

- Technological (product or process) innovators (% of all SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BE, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

- Non-technological (marketing or organisational) innovators (% of all SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BE, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

- Reduced labour costs resulting from process innovations (% of SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

- Reduced use materials and energy resulting from process innovations (% of SMEs)

Availability	2004	2006	2008
NUTS 1	AT, BE, BG, FR, UK	AT, BG, FR, UK	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, IT, NO, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL		

Expenditure-based indicators

These indicators calculate the share of certain types of spending our sales out of total turnover:

- Non-R&D innovation expenditures (% of total turnover)

Availability	2004	2006	2008
NUTS 1	BE, BG, FR	BE, BG	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL, UK		

- Sales of new-to-market products (% of total turnover)

Availability	2004	2006	2008
NUTS 1	BE, BG, FR	BE, BG	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL, UK		

- Sales of new-to-firm products (% of total turnover)

Availability	2004	2006	2008
NUTS 1	BE, BG, FR	BE, BG	AT, BE, BG, FR
NUTS 2	CZ, ES, FI, NO, PL, PT, RO, SI, SK	CZ, ES, FI, GR, HU, PL, PT, RO, SI, SK	CZ, ES, FI, HU, IT, NO, PL, PT, RO, SI, SK, SE
N/A	DE, IE, NL, UK		

Regional CIS data availability is relatively poor as compared to the "non-CIS" indicators:

- Regional CIS data are completely missing for the regions in Germany, Ireland and the Netherlands.
- Expenditure data are not available for the regions in Hungary for 2004, Austria, Italy and Sweden for 2004 and 2006, Greece for 2004 and 2008, France for 2006 and the UK for all years.
- Share data are not available for the regions in, Hungary for 2004, Sweden for 2004 and 2006, Greece for 2004 and 2008 and the UK for 2008.

In the following section we will discuss the procedure used for imputing missing data.

3. Methodology

For the selection of regions and indicators a significant amount of data is missing, in particular for the regions of Germany, Ireland and Netherlands for which no regional CIS data are available. For the analysis in the RIS report two options are available:

- Exclude all regions for which too many data are missing
- Impute the missing data by statistically estimating these missing values using the data which are available

Choosing the first option would significantly reduce the value of the RIS as not all European regions would be covered, including regions in 3 of the most innovative countries. We therefore estimate all missing data using the same procedure as used in the RIS 2009.

3.1 Imputation of missing data

Consider a missing value for indicator Y in region R for a given year:

1. Seek for indicator Z the highest correlation with indicator Y (Z can be the same indicator as Y at another time point or a different indicator at any time point).
2. If the correlation between Y and Z is higher than 0.6 and a value is available for indicator Z in region R, THEN impute a value for Y in region R using Excel's FORECAST function; ELSE use the median ratio procedure described hereafter.

The median ratio imputation procedure calculates, for each indicator and each year, the ratio between the score of a region R and that of the country C to which the region belongs. Then the median across all the indicators and all time points is computed. Then, if indicator Y in region R is missing, it can be estimated by assuming that the same median ratio between R and C applies also to indicator Y. This procedure implies that a score already exists for indicator Y in country C. If this is not the case, the same procedure can be applied one level up in the hierarchy, i.e. between country C and the EU27 aggregate. Data for the EU27 aggregate are available for all indicators.

Most of the imputations have been made via the median ratio procedure: the FORECAST function allows the imputation of only a small percentage of the missing values due to the existence of many regions with missing values for pairs of correlated indicators.

These imputed values are estimates and as such are affected by uncertainty. The uncertainty analysis in the RIS 2009 methodology report[2] showed that "it is not possible to attribute ranks to individual regions due to consistent overlaps between their uncertainty intervals". The RIS 2009 therefore did "not construct a ranking of regions" but instead analysed "groupings of regions based on their overall level of innovation performance". For the same reasons the RIS 2012 also does not provide individual rank results.

3.2 Normalisation of the indicators

Most of the indicators are fractional indicators with values between 0% and 100%. Some indicators are unbound indicators, where values are not limited to an upper threshold. These indicators can have skewed data distributions (where most regions show low performance levels and a few regions show exceptionally high

[2] Cf. footnote 1.

performance levels). For all indicators data will be transformed using a square root transformation with power *N* if the degree of skewness of the raw data exceeds 0.5 such that the skewness of the transformed data is below 0.5 (none of the imputed data are included in this process):

$$\tilde{X}_r = \sqrt[N]{X_r}$$

Table 3 summarizes the degree of skewness before and after the transformation and the power *N* used in the transformation. The data are then normalized using the min-max procedure where the transformed score is first subtracted with the minimum score over all regions in 2004, 2006 and 2008 and then divided by the difference between the maximum and minimum scores over all regions in 2004, 2006 and 2008, where the maximum normalised score is equal to 1 and the minimum normalised score is equal to 0:

$$\hat{X}_r = \frac{\tilde{X}_r - MIN(\forall_r \tilde{X}_r)}{MAX(\forall_r \tilde{X}_r) - MIN(\forall_r \tilde{X}_r)}$$

For each year a composite regional innovation index (RII) is calculated as the unweighted average of the re-scaled scores for all indicators.

Table 3: Degree of skewness and transformation

	Degree of skewness before transformation	Power used in transformation	Degree of skewness after transformation
1.1.2 Population having completed tertiary education	0.262	No transformation	
1.3.1 R&D expenditure in the public sector	1.004	½	-0.039
x.x.x Households with broadband access	0.590	½	-0.016
x.x.x Attitude to entrepreneurship	0.271	No transformation	
2.1.1 R&D expenditure in the business sector	1.708	¼	0.111
2.1.2 Non-R&D innovation expenditure	1.458	¼	-0.032
x.x.x Capital stock	-0.071	No transformation	
2.2.1 SMEs innovating in-house	0.034	No transformation	
2.2.2 Innovative SMEs collaborating with others	0.360	No transformation	
2.2.3 Public-private co-publications	6.131	¼	-0.072
2.3.1 EPO patents	2.171	¼	0.234
3.1.1 Product and/or process innovators	0.174	No transformation	
3.1.2 Marketing and/or organisational innovators	0.784	½	0.243
3.1.3 Resource efficiency innovators: Labour/Energy	0.903	½	0.298
3.2.1a Employment in medium-high and high-tech manufacturing	0.617	½	0.009
3.2.1b Employment in knowledge-intensive services	0.896	½	0.145
3.2.4a New-to-market sales	0.362	No transformation	
3.2.4b New-to-firm sales	0.785	½	0.224

4. Robustness test of clustering methods

The RIS 2009 included a grouping of the European regions that was developed through the application of cluster analysis. A clustering procedure was carried out on the regional scoreboard results in order to cluster regions into five groups according to the similarity of their innovation systems. The five categories were: high performers, medium-high performers, average performers, medium-low performers and low performers[3].

The RIS 2012 will also include a grouping achieved through cluster analysis. The purpose of this report is to take a preliminary set of the RIS data for 2011 and to apply various 'robustness' tests in order to understand the effect of using differing cluster algorithms on the resulting groupings. The purpose is to understand the extent to which groupings achieved through cluster analysis are likely to represent genuine real-world differences between the regional innovation systems of the regions modelled or, alternatively, whether the groupings are highly sensitive to the statistical technique employed. Beyond this, the purpose is to understand what the best approach to clustering the RIS indicators might be.

Why are robustness tests necessary?

All statistical techniques require the application of judgement and interpretation. Judgement is required to decide which technique to use, which variables to include, what relationships to model and how to interpret results, regardless of the particular statistical technique that is employed. This is true whether techniques involve the estimation of model parameters and the application of tests for statistical significance, or whether, as in cluster analysis, they do not.

Because cluster analysis does not try to fit an empirical representation of a theoretical model to data, the method imposes fewer assumptions on the data (for example, about the functional form of the model or the statistical distribution of stochastic errors). The key element of judgement lies in the choice of indicators to include, on the basis of which cases are to be treated as being similar or dissimilar to each other. A consequence of imposing fewer assumptions is that the conclusions that can be drawn are weaker: the results do not confirm or reject a particular theory, and there are no tests of statistical significance of results to be applied. But the method is not vulnerable to the risk that those underlying assumptions of functional form and statistical distribution are incorrect. A particular issue in the application of cluster analysis is to determine how many clusters can reasonably be distinguished in a given data set, and this is one area where some element of judgement typically enters the analysis. One way of addressing this systematically, which we apply as one of our tests in this study, is to make use of the error-sum-of-squares statistic that can be calculated as an output from Ward's method of clustering.

A related issue in the application of cluster analysis is the choice of indicators to include. It may be that adding an additional indicator causes a marked change to the resulting set of clusters that are distinguished, in which case the user has to decide what interpretation to place on that finding and whether the selection of any particular set of clusters can be regarded as a robust outcome of the analysis. The same problem arises with parametric statistical methods, but there the method itself provides a clear guide as to the extent to whether the results depend on the inclusion of a particular indicator, in the form of tests of significance. In cluster analysis, the usual approach is simply to carry out repeated analysis with different sets of indicators and to look for interesting differences in outcomes.

[3] See p.10 of the Regional Innovation Scoreboard 2009

4.1 Clustering methods

There are two main types of cluster analysis: hierarchical cluster analysis and non-hierarchical cluster analysis. The distinction between these is explained in this chapter and it is shown that one way to achieve more robust results is to use them in combination. Three different types of cluster analysis, Ward's method, Average 'between-groups' linkages and k-means, are explained as these are the techniques used in the subsequent robustness analysis.

4.1.1 Hierarchical cluster analysis

Hierarchical clustering procedures start by assuming that each and every case represents a 'category' in its own right. In the first iteration of the clustering algorithm the two cases that are most similar (or 'closest' to each other) are then merged to form a bigger cluster with the degree of similarity (or 'distance') measured by the amount of information lost by the merger.

This is why these clustering methods are known as 'hierarchical' – they assume that every case is a cluster in its own right and then hierarchically merge cases or clusters to produce a smaller and smaller number of clusters as the algorithm proceeds. The difficulty is to know when to halt the procedure. In other words, a decision has to be taken as to when the 'true' number of clusters embedded in the data has emerged and the procedure has to be halted at that point or it will continue merging clusters until only one cluster containing all cases remains. As noted in the previous chapter, the role required here for judgement by the user (to decide when to halt the procedure because the 'true' number of clusters has emerged) is a reason for applying tests of robustness in order to check whether such judgements make much difference to the key findings.

By far the most commonly employed hierarchical clustering method is Ward's method, which was used for the RIS 2009 and is described below.

Ward's method

The dissimilarity between clusters is measured on the basis of the information that would be lost by merging them into one. In each iteration, pairs of clusters are considered one at a time until all pairs have been considered and then the pair that can be merged with the least loss of information is selected for merging. The 'error-sum-of-squares' (or 'cluster coefficient) is the measure of information loss produced at each point of merger. This measure increases in small increments for the majority of the clustering procedure as relatively similar cases are merged. The point at which there is a jump in the error-sum-of-squares is the point at which two quite dissimilar clusters of cases have been merged. The 'true' number of clusters emergent from the data can therefore be taken as the number of clusters existing just prior to this large jump in the error-sum-of-squares. This is therefore one method of testing the robustness of any grouping. Ward's method of cluster analysis can be used to ascertain the 'true' number of emergent clusters embedded in the data and then alternative methods of clustering can be applied to see whether the number of clusters (and the cases within each cluster) remains.

Average 'between-groups' linkage

Clusters are considered in pairs. The dissimilarity between two clusters is calculated as the average distance between all pairs of cases within the two clusters. In each iteration, the two clusters are merged in which this average distance is lowest.

4.1.2 Squared Euclidean Distance

In the above descriptions reference has been made to 'distance' between, for example, cases. The 'standard' way to measure distance when using hierarchical cluster analysis is 'squared Euclidean distance'. Squared Euclidean Distance is based on an extension of Pythagoras' theorem that allows for the calculation of geometric distance between two points in multi-dimensional space. The measure of distance is squared in order to lend greater weight to distances that are further apart because, as previously stated, cluster analysis seeks to maximise variability between clusters and minimise variability within clusters.

4.1.3 Non-hierarchical (k-means) cluster analysis

Non-hierarchical cluster analysis refers to what is known as k-means cluster analysis. The major distinction between the hierarchical method and this method is that under the hierarchical method the number of clusters is emergent and can be judged for example by using the error-sum-of-squares output. Non-hierarchical methods require the user to specify the desired number of clusters in advance.

Non-hierarchical cluster analysis is often used when the dataset is large. It is sometimes considered superior or more accurate than hierarchical clustering because it allows cases to move between clusters iteratively until a best-fit is found. In contrast, once a case is assigned to a cluster in hierarchical cluster analysis it remains there and cannot move.

Clusters are formed in non-hierarchical cluster analysis by assigning the case to the cluster with the nearest mean. The cases are somewhat arbitrarily assigned initially, then the process of comparing case means with cluster means begins and cases are iteratively reassigned to new clusters, whose means change as cases are re-assigned to them. The clustering algorithm stops when all clusters are stable – when no case would be nearer to the mean if it was moved to another cluster.

Because of the differing methods by which clusters are formed using the non-hierarchical approach described above, and the fact it is sometimes considered more accurate, a useful approach can be firstly to carry out hierarchical analysis using, say, Ward's method, and then to follow this up with k-means cluster analysis by using the number of clusters determined using the hierarchical approach. It is then possible to see if that number of clusters continues to produce similar results using the non-hierarchical methods – for example, do the cases end up in the same clusters? This is a useful way of using the purportedly more accurate k-means approach without having arbitrarily to decide the number of clusters a priori. We apply this procedure below.

4.2 Robustness tests

The robustness test to be carried out on the currently available RIS data for 2011 has taken the following form:

1. Determine the number of clusters in the data through the application of hierarchical cluster analysis employing Ward's method and Squared Euclidean Distance. Document the allocation of cases to clusters for subsequent comparison.
2. Carry out one further type of hierarchical cluster analysis (Average 'between-groups' linkages) with Squared Euclidean Distance, as well as one example of non-hierarchical cluster analysis (k-means), using the number of clusters established in 1) as an input, and compare the allocation of cases (regions) to that achieved in 1).

If most cases (regions) grouped together under 1) are also grouped together using the alternative techniques examined in 2) then the clustering achieved under 1) is considered robust and representative of genuine differences between groups of regions. However, if cluster membership is not consistent when different techniques are applied, then the outcome of 1) is not considered robust and hierarchical cluster analysis using Ward's method and Squared Euclidean Distance, as employed for the 2009 RIS, cannot be considered to produce particularly reliable results.

More generally, we are checking to see the extent to which distinctive clusters that have an interesting theoretical interpretation are produced. The first task is therefore to carry out a standard hierarchical clustering employing Ward's method and using Squared Euclidean Distance.

The following indicators have been included in the robustness test and for all indicators 2008 data have been used:

- Population with tertiary education per 100 population aged 25-64
- Public R&D expenditures (% of GDP)
- Households with broadband access
- Attitude to entrepreneurship
- Business R&D expenditures (% of GDP)
- Non-R&D innovation expenditures (% of total turnover)
- SMEs innovating in-house (% of all SMEs)
- Innovative SMEs collaborating with others (% of all SMEs)
- EPO patents per million population
- Capital stock per million population
- Public-private co-publications per million population
- Technological (product or process) innovators (% of all SMEs)
- Non-technological (marketing or organisational) innovators (% of all SMEs)
- Resource efficiency innovators
- Employment in knowledge-intensive services (% of total workforce)
- Employment in medium-high and high-tech manufacturing (% of total workforce)
- Sales of new-to-market products (% of total turnover)
- Sales of new-to-firm products (% of total turnover)

Households with broadband access and Resource efficiency innovators are included as these indicators were also used in the RIS 2009. Attitude to entrepreneurship and Capital stock per million population are included as these measure relevant aspects of innovation. These indicators were not included in the RIS 2009 as the data were not available at that time. Public-private co-publications were not available in the RIS 2009 but the indicator is included in IUS. For EPO patents we use per million population in the denominator similar as in the RIS 2009, but in IUS the indicator is defined per million GDP.

4.2.1 Hierarchical cluster analysis using Ward's method

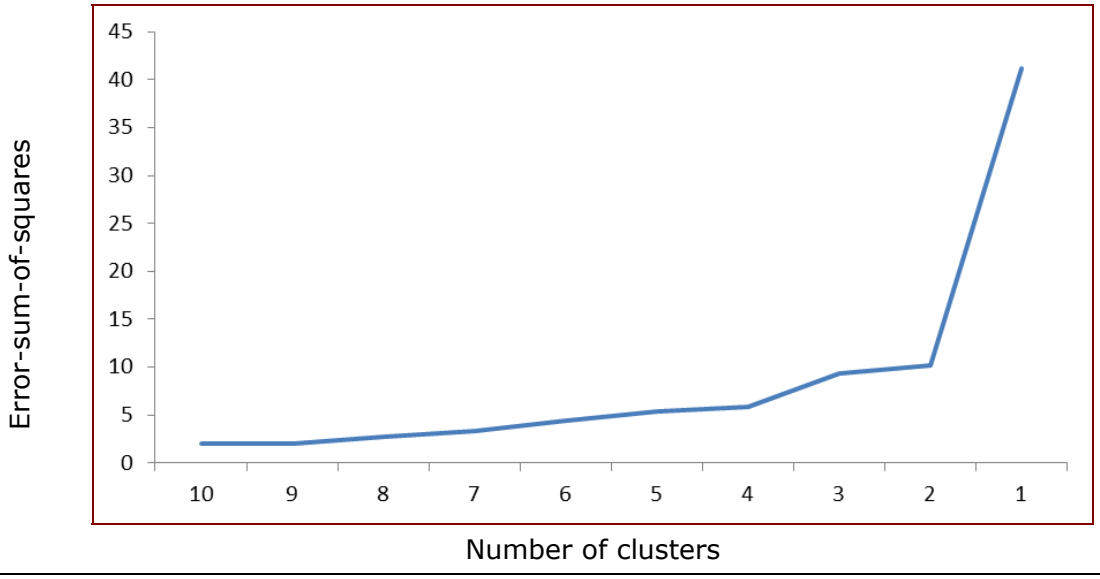
As stated in the previous chapter, one of the advantages of using Ward's method to carry out cluster analysis is that it produces an output showing the error-sum-of-squares upon every iterative merger, and the pattern of increases in this statistic can be used to inform judgement about the number of clusters that should be identified in the data.

Table 4 below shows the increase in the error-sum-of-squares for each of the last ten steps of the hierarchical clustering procedure using Ward's method carried out on the currently available RIS data[4]. Figure 1 illustrates the change.

Table 4: Change in error-sum-of-squares

No. of clusters	Error-sum-of squares	Increase in error-sum-of -squares
10	44.856	1.987
9	46.851	1.995
8	49.540	2.689
7	52.861	3.321
6	57.340	4.479
5	62.715	5.375
4	68.550	5.835
3	77.929	9.379
2	88.147	10.218
1	129.268	41.121

Figure 1: Increase in error-sum-of-squares



As is clear from Figure 1, there is a sharp inflection when only two clusters remain and these are merged into one. However, this is to be expected as considerable information is lost when all cases are merged into one single cluster and it is not a very useful result to say there is just one cluster representing all cases. The first inflection which occurs is when four clusters are merged into three. As can be seen in Table 4, while the increase in the error-sum-of-squares remains quite small even during this step, it increases somewhat more at this point compared to the increase in the previous steps. This suggests that we should consider there to be four distinct clusters emergent from the data.

Table 5 shows the mean value for each indicator for each cluster using hierarchical cluster analysis with Ward's method and Squared Euclidean Distance. Since cluster analysis maximises the variability between clusters, the examination

[4] In all instances the indicators are those shown in Table 5.

of the cluster mean for each indicator helps us to give an interpretative characterisation to each cluster.

The first column of the table in Annex A shows which cases (regions) fall into which cluster under the four-cluster solution using hierarchical cluster analysis with Ward's method shown above. The subsequent columns show the allocation of cases to clusters under the two subsequent alternative clustering methods for comparison purposes.

In Table 5 the highest mean score on each indicator has been highlighted in bold and the lowest mean score has been highlighted in italic. The Ward's method results distinguish two clusters with regions that are low (cluster 3) and high (cluster 4) performing on most of the indicators, together with two other clusters both of which include some indicators with the highest mean values. In this application of clustering, where all of the indicators are intended to represent positive influences on, or measures of, innovative activity, the fact that the method has identified two clusters in which most indicators are similarly high or low is a helpful result. The results for the other two groups are less easily interpreted, but further analysis might show some other interesting dimension of difference (for example, on the basis of size of firms).

Table 5: Comparison of cluster means under Ward's method

	Mean for cluster 1	**Mean for cluster 2**	**Mean for cluster 3**	**Mean for cluster 4**
Population with tertiary education	**0.62**	0.42	*0.30*	0.45
Public R&D expenditures	0.57	0.50	*0.35*	**0.61**
Households with broadband access	**0.89**	0.71	*0.62*	0.85
Attitude to entrepreneurship	0.50	**0.54**	0.44	*0.36*
Business R&D expenditures	**0.66**	0.50	*0.26*	0.65
Non-R&D innovation expenditures	0.64	0.68	*0.63*	**0.95**
SMEs innovating in-house	0.57	0.49	*0.21*	**0.60**
Innovative SMEs collaborating with others	**0.57**	0.36	*0.19*	0.47
EPO patents per million population	0.62	0.44	*0.21*	**0.78**
Capital stock per million population	0.65	0.45	*0.22*	**0.72**
Public-private co-publications	**0.59**	0.35	*0.15*	0.52
Technological (product or process) innovators	0.51	0.49	*0.22*	**0.84**
Non-technological (marketing or organisational) innovators	0.46	0.50	*0.25*	**0.89**
Resource efficiency innovators	0.37	0.45	*0.22*	**0.90**
Employment in knowledge-intensive services	**0.66**	0.53	*0.36*	0.64
Employment in medium-high and high-tech manufacturing	0.39	0.47	*0.33*	**0.67**
Sales of new-to-market products	0.46	**0.50**	0.33	*0.22*
Sales of new-to-firm products	0.50	0.62	*0.47*	**0.91**
No. of cases (regions) in cluster	44	74	63	11

In sum, the Ward's method approach produces two clusters that are quite clearly distinct. The case for distinguishing the other two is less clear, and, in fact, if further merging to three clusters is allowed, most of the regions in clusters 1 and 2 end up in the same cluster.

4.2.2 Hierarchical cluster analysis using Average 'between-groups' linkage and Squared Euclidean Distance

The number of clusters implied by Ward's method is used as an input to the clustering process here. The idea is to try to understand whether an alternative cluster technique produces even more distinct clusters and whether the clusters are similar to those produced using Ward's method in terms of the cluster membership of regions (as shown in Annex A).

As seen in Table 6, the Average 'between-groups' linkage clustering algorithm has produced fairly distinctive clusters too. Cluster 3 is clearly composed of the regions that are high-level innovators. Cluster 3 has the highest mean, represented by scores highlighted in bold, on a higher number (thirteen) indicators than its equivalent under Ward's method (cluster 4), but on two of these thirteen it ties for highest with cluster 1 (households with broadband access and attitude to entrepreneurship). In terms of the distinctiveness of the cluster representing higher-level innovating regions this technique has therefore produced similar results to that achieved using Ward's method. In terms of the cluster that clearly represents the lowest innovating regions (cluster 2) the result is somewhat less distinctive than under Ward's method where cluster 3 represented the lowest score on all indicators apart from one. Here cluster 2 clearly represents low-level innovating regions but the lowest mean scores are less concentrated in this cluster as some are scattered across the other clusters.

Table 6: Comparison of cluster means under Average 'between-groups' linkage method

	Mean for cluster 1	Mean for cluster 2	Mean for cluster 3	Mean for cluster 4
Population with tertiary education	**0.50**	0.36	0.47	*0.16*
Public R&D expenditures	0.52	*0.37*	**0.62**	0.46
Households with broadband access	**0.79**	*0.64*	**0.79**	0.68
Attitude to entrepreneurship	**0.52**	0.46	**0.52**	*0.35*
Business R&D expenditures	**0.59**	*0.29*	0.58	0.37
Non-R&D innovation expenditures	0.68	0.61	**0.92**	*0.60*
SMEs innovating in-house	0.54	*0.21*	0.55	**0.78**
Innovative SMEs collaborating with others	0.46	*0.20*	0.42	**0.52**
EPO patents per million population	0.54	*0.23*	**0.70**	*0.23*
Capital stock per million population	0.54	*0.26*	**0.64**	0.31
Public-private co-publications	0.47	*0.18*	**0.48**	0.21
Technological (product or process) innovators	0.50	*0.23*	**0.76**	0.75
Non-technological (marketing or organisational) innovators	0.47	*0.27*	**0.82**	0.58
Resource efficiency innovators	0.39	*0.23*	**0.83**	0.70
Employment in knowledge-intensive services	0.59	*0.39*	**0.60**	0.42
Employment in medium-high and high-tech manufacturing	0.46	*0.33*	**0.60**	0.28
Sales of new-to-market products	**0.51**	0.34	*0.22*	0.46
Sales of new-to-firm products	0.56	*0.49*	**0.86**	0.66
No. of cases (regions) in cluster	94	74	17	7

A further question to ask is whether cases (regions) are allocated to similar clusters under this method compared to in Ward's method. For example, are cases in the cluster representing high-level innovating regions (cluster 3) equivalent to those in the equivalent cluster (cluster 4) under Ward's method?

Annex A shows that all seven regions allocated to cluster 4 under Ward's method have been allocated to cluster 3, the equivalent 'high-level' innovating cluster, under this method – but the cluster is expanded and represents seventeen instead of seven cases under this alternative method as it includes some regions allocated to cluster 2 under Ward's method.

The results are similar to Ward's method and therefore suggest there is some 'robustness' to using cluster analysis to group the RIS regions as results can be replicated using different methods. However, the results using the above method are not as distinctive (with respect to interpretation of innovative performance) as those achieved using Ward's method.

4.2.3 Non-hierarchical k-means cluster analysis

Table 7 shows the results of a non-hierarchical cluster analysis carried out using the number of clusters (four) implied by the previously implemented Ward's method as an input. Annex A again compares the allocation of cases under this method to that under the previous methods. The scores highlighted in bold show the highest among the four clusters and those highlighted in italic the lowest.

The k-means clustering algorithm has produced clusters which have a similar distinctiveness to that achieved using Ward's method. There is one cluster, cluster 2, which clearly groups together the most innovative regions in Europe. On twelve of the eighteen indicators cluster 2 has the highest mean.

In fact, cluster 2 represents an expanded equivalent to cluster 4 under the Ward's method analysis. All of the cases that were in cluster 4 under Ward's method are now in cluster 2 under the k-means method. The k-means cluster analysis' iterative approaches which allows for cases to move between clusters until a best fit is found has therefore resulted in a somewhat more equal distribution of regions between the clusters.

Table 7: Comparison of cluster means under k-means method

	Mean for cluster 1	Mean for cluster 2	Mean for cluster 3	Mean for cluster 4
Population with tertiary education	**0.61**	0.49	*0.28*	0.33
Public R&D expenditures	0.56	**0.64**	0.43	*0.36*
Households with broadband access	0.86	**0.81**	0.65	*0.63*
Attitude to entrepreneurship	0.51	**0.52**	0.48	*0.47*
Business R&D expenditures	**0.62**	**0.62**	0.44	*0.27*
Non-R&D innovation expenditures	0.65	**0.91**	0.67	*0.61*
SMEs innovating in-house	0.54	**0.61**	0.51	*0.19*
Innovative SMEs collaborating with others	**0.52**	0.46	0.31	*0.19*
EPO patents per million population	0.58	**0.70**	0.39	*0.21*
Capital stock per million population	0.60	**0.65**	0.41	*0.21*
Public-private co-publications	**0.52**	0.51	0.29	*0.16*
Technological (product or process) innovators	0.49	**0.77**	0.51	*0.20*
Non-technological (marketing or organisational) innovators	0.44	**0.81**	0.51	*0.25*
Resource efficiency innovators	0.39	**0.78**	0.42	*0.21*
Employment in knowledge-intensive services	**0.63**	0.62	0.49	*0.36*
Employment in medium-high and high-tech manufacturing	0.41	**0.59**	0.47	*0.33*
Sales of new-to-market products	0.46	*0.28*	**0.57**	0.31
Sales of new-to-firm products	0.51	**0.87**	0.63	*0.47*
No. of cases (regions) in cluster	65	19	45	63

Cluster 4 is clearly composed of regions that are 'low-level innovators' as the cluster mean is the lowest on all but two indicators. Cluster 1 is then composed of medium-high innovating regions as its mean is highest on quite a few indicators, though not as many as cluster 2. Finally, cluster 3 represents medium-low innovating regions as its mean is the highest on only one indicator and the lowest on only one indicator too.

The results are somewhat similar to those achieved using Ward's method but with a somewhat more equal distribution of regions among clusters.

4.3 Conclusions cluster analysis

The analysis carried out for this report suggests that the different methods used produce relatively similar results in terms of the distinctiveness of clusters. There can be some variability in terms of the allocation of cases to clusters, however. In particular, Average 'between-groups' linkage and k-means methods produced a more expanded high-level innovating group of regions. The limited group of just seven regions under Ward's method are all German, and it may therefore be more interesting for policy reflection to consider the features of the wider group of regions that includes non-German cases.

A combination of Ward's method and k-means may produce the 'best' results

To achieve the best results it may be advisable to employ Ward's method to determine the number of emergent clusters in the data and then to apply k-means cluster analysis to check whether the characteristics of the clusters and the allocation of regions are similar under the two methods, and that the resulting clusters have an interesting interpretation. In the example shown, we have found that they are reasonably similar.

Ward's method has the advantage of producing the cluster coefficient error-sum-of-squares as an output and this can be charted, as shown in Figure 1, in order to determine in a non-arbitrary and emergent fashion a likely number of clusters embedded in the data. However, a disadvantage of the Ward's method approach is that it simply allocates cases to clusters hierarchically and there is no iterative component that allows cases to move between clusters. K-means cluster analysis provides for a more thorough allocation of cases to clusters because it allows for the iterative shifting of cases between clusters. It is likely to be for this reason that the 'high-level' innovating cluster under Ward's method contained just seven cases compared to nineteen under k-means. K-means cluster analysis, however, has the disadvantage that if it is used in isolation the number of clusters embedded in the data has to be decided arbitrarily a priori and this is why we recommend that it is used in combination with Ward's method.

Other methods that may improve results still further include banding of variables before applying Ward's method.

Cluster analysis should be viewed as interpretivistic in nature

A broader issue relates to the purpose for which cluster analysis is used. The purpose of cluster analysis is to provide a grouping of cases for which a narrative explanation can ideally be built up. The usefulness of any grouping should therefore mostly be assessed on the plausibility and usefulness of the qualitative explanation that is attached to it. There are signals that suggest a grouping may be useful as an explanatory device, some of which have been discussed in this report – the main one is that relatively distinctive clusters are produced. There is no such thing as a 'definitive' or 'correct' grouping, but groupings that emerge

with reasonable consistency under the application of different methods or under alternative selections of the set of indicators clearly command greater confidence than that prove to be less stable in this respect. But clusters that are robust to alternative methods and data sets do not necessarily carry a policy-relevant interpretation.

Cluster analysis should be used as part of the RIS component of the Innovation Union Scoreboard with this in mind. Its purpose is not to provide a definitive grouping but to group regions in a way which may provide for a useful qualitative comparison and contrast between the regional innovation systems in question. It is the usefulness of the qualitative comparison and contrast that can be built up using a particular grouping that determines the usefulness of that grouping.

Is it regional or national-level innovation systems that matter?

One question of interest to emerge from this examination is whether the real drivers of innovation performance are regional or whether it is the macro-level national innovation system that really counts. The clustering algorithms in all three cases tend to allocate regions from the same country to the same cluster, which tends to suggest it is the macro, national-level innovation system that is important for innovation rather than anything systemic at the regional level as Commission policy currently assumes. For example, the Polish regions PL11 to PL63 are allocated to the same cluster under each of the three different clustering methods. Most of the German regions are in the same (high-level innovating) cluster regardless of which method is used. This may suggest that a regional-level analysis is not able to add much more to an explanation of European innovation than an analysis of national-level systems, as carried out by the main IUS. But it would not rule out the possibility that within-country variations in innovative activity are an important driver of regional differences in economic performance.

5. Conclusions for RIS 2012

The Innovation Union Scoreboard (IUS) has become a well-established benchmarking tool for monitoring Member States' innovation performance. The IUS uses a range of indicators at the country level to summarize performance in a single number or composite indicator. At the regional level a comparable exercise is not possible due to limited regional data availability. A regional innovation scoreboard was first introduced in 2004 using a small number of indicators and was updated in 2006 when for some indicators unpublished CIS data could be included. In 2008-2009 Member States were asked to share regional CIS data for a larger number of indicators allowing a more elaborate RIS report in 2009 covering 16 indicators, of which 8 used CIS data, for 201 regions.

After the revision of the European Innovation Scoreboard into the Innovation Union Scoreboard in 2010, it is time for another update of the RIS. The RIS 2012 will adopt the IUS methodology at the regional level. The previous sections have shown that only for half the IUS indicators regional data are available. In particular for the group of Enablers data availability is poor.

For the RIS 2012 there are at least three options (cf. Table 8). If one would only use those indicators which are identical to the indicators used in IUS, the RIS 2012 would only cover 8 indicators. If we broaden the RIS indicators and also include those indicators for which close proxies are available the number of indicators included in RIS 2012 would increase to 12. If we would add a few indicators relevant for measuring innovation but which are not included in IUS the number of indicators included in RIS 2012 would increase to 16.

For comparability issues option 2 is the preferred option, i.e. to only include those indicators which are either identical to those used in IUS or which are very close proxies for IUS indicators.

These 12 indicators will be used to construct a regional innovation index (RII) using a methodology similar but different as that used in the IUS. The RII will be calculated as a weighted average of the normalised data of the 12 indicators (cf. Annex B for more details). Similar as in the RIS 2009, the RII data will be used to classify regions in different performance groups. These performance groups will be compared with the 4 performance groups identified in the IUS (i.e. innovation leaders, innovation followers, moderate innovators and modest innovators).

The RII will also be compared with the EU Regional Competitiveness Index[5] to identify possible linkages between regions' competitiveness and innovation performance.

For the indicators capturing capital stocks, entrepreneurial attitudes, resource efficiency innovators and structural funds allocations separate analyses are foreseen for the final RIS report providing more detail on differences between regions' performance.

[5] Annoni, P. and K. Kozovska, EU Regional Competitiveness Index 2010, JRC, EUR 24346 EN (http://bookshop.europa.eu/en/eu-regional-competitiveness-index-2010-pbLBNA24346/)

Table 8: Options for indicators to be included in RIS 2012

Innovation Union Scoreboard	**Option 1** **Identical indicators**	**Option 2** **Identical indicators + proxies**	**Option 3** **Identical indicators + proxies + additional indicators**
	8 indicators	**12 indicators**	**16 indicators**
Human resources			
1.1.1 New doctorate graduates (ISCED 6) per 1000 population aged 25-34	No	No	No
1.1.2 Percentage population aged 30-34 having completed tertiary education	No	Percentage population aged 25-64 having completed tertiary education	Percentage population aged 25-64 having completed tertiary education
1.1.3 Percentage youth aged 20-24 having attained at least upper secondary level education	No	No	No
Open, excellent and attractive research systems			
1.2.1 International scientific co-publications per million population	No	No	No
1.2.2 Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country	No	No	No
1.2.3 Non-EU doctorate students as a % of all doctorate students	No	No	No
Finance and support			
1.3.1 R&D expenditure in the public sector as % of GDP	Identical	Identical	Identical
1.3.2 Venture capital (early stage, expansion and replacement) as % of GDP	No	No	No
--	--	--	Households with broadband access
--	--	--	Structural fund allocations on core RTDI activities per mln population Structural fund allocations on business innovation per mln population
Firm investments			
2.1.1 R&D expenditure in the business sector as % of GDP	Identical	Identical	Identical
2.1.2 Non-R&D innovation expenditures as % of turnover	No	Similar (only for SMEs)	Similar (only for SMEs)
--	--	--	Capital stock data per mln population
Linkages & entrepreneurship			
2.2.1 SMEs innovating in-house as % of SMEs	Identical	Identical	Identical
2.2.2 Innovative SMEs collaborating with others as % of SMEs	Identical	Identical	Identical
2.2.3 Public-private co-publications per million population	Identical	Identical	Identical

Innovation Union Scoreboard	Option 1 Identical indicators	Option 2 Identical indicators + proxies	Option 3 Identical indicators + proxies + additional indicators
Intellectual assets			
2.3.1 PCT patents applications per billion GDP (in PPS€)	Identical	Identical	Identical
2.3.2 PCT patent applications in societal challenges per billion GDP (in PPS€)	No	No	No
2.3.3 Community trademarks per billion GDP (in PPS€)	No	No	No
2.3.4 Community designs per billion GDP (in PPS€)	No	No	No
Innovators			
3.1.1 SMEs introducing product or process innovations as % of SMEs	Identical	Identical	Identical
3.1.2 SMEs introducing marketing or organisational innovations as % of SMEs	Identical	Identical	Identical
3.1.3 High-growth innovative firms	N/A	N/A	N/A
--	--	--	Resource efficiency innovators
Economic effects			
3.2.1 Employment in knowledge-intensive activities (manufacturing and services) as % of total employment	No	Employment in knowledge-intensive services + Employment in medium-high/high-tech manufacturing as % of total workforce	Employment in knowledge-intensive services + Employment in medium-high/high-tech manufacturing as % of total workforce
3.2.2 Medium and high-tech product exports as % total product exports	No	No	No
3.2.3 Knowledge-intensive services exports as % total service exports	No	No	No
3.2.4 Sales of new to market and new to firm innovations as % of turnover	No	Similar (only for SMEs)	Similar (only for SMEs)
3.2.5 License and patent revenues from abroad as % of GDP	No	No	No

Annex A: Allocation of cases to clusters under different clustering methods

Region code	Region name	Ward's method	Average between-groups linkages	K-means
AT1	Ostösterreich	1	1	1
AT2	Südösterreich	1	1	1
AT3	Westösterreich	1	1	1
BE1	Région de Bruxelles-Capitale / Brussels	1	1	1
BE2	Vlaams Gewest	1	1	1
BE3	Région Wallonne	2	1	1
BG3	Severna i iztochna Bulgaria	3	2	4
BG4	Yugozapadna i yuzhna tsentralna Bulgaria	3	2	4
CH01	Région lémanique	1	1	1
CH02	Espace Mittelland	1	1	1
CH03	Nordwestschweiz	1	1	2
CH04	Zürich	1	1	2
CH05	Ostschweiz	1	1	1
CH06	Zentralschweiz	1	1	1
CH07	Ticino	1	1	1
CZ01	Praha	2	1	1
CZ02	Střední Cechy	2	1	3
CZ03	Jihozápad	2	1	3
CZ04	Severozápad	2	1	3
CZ05	Severovýchod	2	1	3
CZ06	Jihovýchod	2	1	3
CZ07	Střední Morava	2	1	3
CZ08	Moravskoslezsko	2	1	3
DE1	Baden-Württemberg	4	3	2
DE2	Bayern	4	3	2
DE3	Berlin	4	3	2
DE4	Brandenburg	2	3	1
DE5	Bremen	4	3	2
DE6	Hamburg	4	3	2
DE7	Hessen	4	3	2
DE8	Mecklenburg-Vorpommern	2	3	1
DE9	Niedersachsen	4	3	2
DEA	Nordrhein-Westfalen	4	3	2
DEB	Rheinland-Pfalz	4	3	2
DEC	Saarland	4	3	2
DED	Sachsen	2	3	1
DEE	Sachsen-Anhalt	2	3	1
DEF	Schleswig-Holstein	4	3	2
DEG	Thüringen	2	3	3
ES11	Galicia	2	2	3
ES12	Principado de Asturias	2	2	1
ES13	Cantabria	2	2	1
ES21	País Vasco	2	1	1
ES22	Comunidad Foral de Navarra	2	1	1
ES23	La Rioja	2	2	3
ES24	Aragón	2	2	3
ES3	Comunidad de Madrid	2	1	1
ES41	Castilla y León	2	2	3
ES42	Castilla-la Mancha	3	2	4
ES43	Extremadura	3	2	4
ES51	Cataluña	2	1	3
ES52	Comunidad Valenciana	3	2	4
ES53	Illes Balears	3	2	4
ES61	Andalucía	3	2	4
ES62	Región de Murcia	3	2	4
ES63	Ciudad Autónoma de Ceuta (ES)	3	2	4
ES64	Ciudad Autónoma de Melilla (ES)	3	2	4
ES7	Canarias (ES)	3	2	4

Region code	Region name	Ward's method	Average between-groups linkages	K-means
FI13	Itä-Suomi	2	1	3
FI18	Etelä-Suomi	1	1	1
FI19	Länsi-Suomi	1	1	1
FI1A	Pohjois-Suomi	1	1	1
FR1	Île de France	1	1	1
FR2	Bassin Parisien	2	1	3
FR3	Nord - Pas-de-Calais	2	1	3
FR4	Est (FR)	2	1	3
FR5	Ouest (FR)	2	1	3
FR6	Sud-Ouest (FR)	2	1	3
FR7	Centre-Est (FR)	1	1	1
FR8	Méditerranée	2	1	3
GR11	Anatoliki Makedonia, Thraki	3	2	4
GR12	Kentriki Makedonia	3	2	4
GR13	Dytiki Makedonia	3	2	4
GR14	Thessalia	3	2	3
GR21	Ipeiros	3	2	4
GR22	Ionia Nisia	3	2	4
GR23	Dytiki Ellada	3	2	4
GR24	Sterea Ellada	3	2	4
GR25	Peloponnisos	3	2	4
GR3	Attiki	2	3	3
GR41	Voreio Aigaio	3	2	4
GR42	Notio Aigaio	3	2	4
GR43	Kriti	3	2	4
HR01	Sjeverozapadna Hrvatska	2	1	3
HR02	Sredisnja i Istocna (Panonska) Hrvatska	3	2	4
HR03	Jadranska Hrvatska	3	2	4
HU1	Közép-Magyarország	2	2	3
HU21	Közép-Dunántúl	3	2	3
HU22	Nyugat-Dunántúl	3	2	4
HU23	Dél-Dunántúl	3	2	4
HU31	Észak-Magyarország	3	2	4
HU32	Észak-Alföld	3	2	4
HU33	Dél-Alföld	3	2	4
IE01	Border, Midland and Western	2	1	3
IE02	Southern and Eastern	2	1	3
ITC1	Piemonte	2	1	3
ITC2	Valle d'Aosta/Vallée d'Aoste	2	1	1
ITC3	Liguria	2	1	1
ITC4	Lombardia	2	1	3
ITD1	Provincia Autonoma Bolzano/Bozen	2	1	3
ITD2	Provincia Autonoma Trento	2	1	1
ITD3	Veneto	2	1	1
ITD4	Friuli-Venezia Giulia	2	1	1
ITD5	Emilia-Romagna	2	1	1
ITE1	Toscana	2	1	1
ITE2	Umbria	2	1	1
ITE3	Marche	2	1	1
ITE4	Lazio	2	1	1
ITF1	Abruzzo	2	1	4
ITF2	Molise	3	2	4
ITF3	Campania	2	1	3
ITF4	Puglia	3	2	4
ITF5	Basilicata	3	2	3
ITF6	Calabria	3	2	3
ITG1	Sicilia	3	2	4
ITG2	Sardegna	3	2	4
NL11	Groningen	1	1	1

Region code	Region name	Ward's method	Average between-groups linkages	K-means
NL12	Friesland (NL)	1	1	1
NL13	Drenthe	1	1	1
NL21	Overijssel	1	1	1
NL22	Gelderland	1	1	1
NL23	Flevoland	1	1	1
NL31	Utrecht	1	1	1
NL32	Noord-Holland	1	1	1
NL33	Zuid-Holland	1	1	1
NL34	Zeeland	1	1	1
NL41	Noord-Brabant	1	1	1
NL42	Limburg (NL)	1	1	1
NO01	Oslo og Akershus	1	1	1
NO02	Hedmark og Oppland	3	2	3
NO03	Sør-Østlandet	1	1	1
NO04	Agder og Rogaland	1	1	1
NO05	Vestlandet	1	1	1
NO06	Trøndelag	1	1	1
NO07	Nord-Norge	1	2	1
PL11	Lódzkie	3	2	3
PL12	Mazowieckie	2	2	4
PL21	Malopolskie	3	2	4
PL22	Slaskie	3	2	4
PL31	Lubelskie	3	2	4
PL32	Podkarpackie	3	2	4
PL33	Swietokrzyskie	3	2	4
PL34	Podlaskie	3	2	4
PL41	Wielkopolskie	3	2	4
PL42	Zachodniopomorskie	3	2	4
PL43	Lubuskie	3	2	4
PL51	Dolnoslaskie	3	2	4
PL52	Opolskie	3	2	4
PL61	Kujawsko-Pomorskie	3	2	4
PL62	Warminsko-Mazurskie	3	2	4
PL63	Pomorskie	3	2	4
PT11	Norte	2	4	3
PT15	Algarve	2	4	3
PT16	Centro (PT)	2	4	3
PT17	Lisboa	2	4	1
PT18	Alentejo	2	4	4
PT2	Região Autónoma dos Açores (PT)	2	4	4
PT3	Região Autónoma da Madeira (PT)	2	4	4
RO11	Nord-Vest	3	2	4
RO12	Centru	3	2	4
RO21	Nord-Est	3	2	4
RO22	Sud-Est	3	1	4
RO31	Sud-Muntenia	3	2	4
RO32	Bucuresti-Ilfo	2	2	1
RO41	Sud-Vest Oltenia	3	2	4
RO42	Vest	3	2	4
SE11	Stockholm	1	1	1
SE12	Östra Mellansverige	1	1	1
SE21	Småland med öarna	2	1	2
SE22	Sydsverige	1	1	1
SE23	Västsverige	1	1	1
SE31	Norra Mellansverige	1	1	1
SE32	Mellersta Norrland	1	1	1
SE33	Övre Norrland	2	1	2
SI01	Vzhodna Slovenija	2	1	2
SI02	Zahodna Slovenija	2	1	2
SK01	Bratislavský kraj	2	2	2

Region code	Region name	Ward's method	Average between-groups linkages	K-means
SK02	Západné Slovensko	3	2	4
SK03	Stredné Slovensko	3	2	4
SK04	Východné Slovensko	3	2	4
UKC	North East (UK)	2	1	3
UKD	North West (UK)	2	1	3
UKE	Yorkshire and The Humber	2	1	3
UKF	East Midlands (UK)	2	1	3
UKG	West Midlands (UK)	2	1	3
UKH	East of England	1	1	1
UKI	London	1	1	1
UKJ	South East (UK)	1	1	4
UKK	South West (UK)	2	1	3
UKL	Wales	2	1	3
UKM	Scotland	2	1	1
UKN	Northern Ireland (UK)	2	2	2

Annex B: Calculation of Regional Innovation Index

The regional innovation index will be calculated as a weighted average of the 12 indicators for which regional data are available. The approach resembles a mix of the methodology used in the RIS 2009 and the Innovation Union Scoreboard (IUS) 2011. In the RIS 2009 a weighting schedule was used which reflected the overall weights of Enablers, Firm activities and Outputs and the overall weights of the CIS indicators in the European Innovation Scoreboard (EIS) 2009. Applying a similar weighting scheme to the RIS 2012 would give the indicator weights as shown in Table B.1.

Table B.1: Indicator weights using RIS 2009 methodology

	Weight in Enablers			**Weight of Enablers in IUS**	**Weight of indicator in RIS**
1.1.2 Percentage population aged 25-64 having completed tertiary education	1/2			8/24	16.67%
1.3.1 R&D expenditure in the public sector as % of GDP	1/2			8/24	16.67%
	Weight of non-CIS indicators in Firm activities	**Weight of indicator in non-CIS indicators**	**Weight in Firm activities**	**Weight of Firm activities in IUS**	**Weight of indicator in RIS**
2.1.1 R&D expenditure in the business sector as % of GDP	2/3	1/3	2/9	9/24	8.33%
2.2.3 Public-private co-publications per million population	2/3	1/3	2/9	9/24	8.33%
2.3.1 PCT patents applications per billion GDP (in PPS€)	2/3	1/3	2/9	9/24	8.33%
	Weight of CIS indicators in Firm activities	**Weight of indicator in CIS indicators**			
2.1.2 Non-R&D innovation expenditures as % of turnover	1/3	1/3	1/9	9/24	4.17%
2.2.1 SMEs innovating in-house as % of SMEs	1/3	1/3	1/9	9/24	4.17%
2.2.2 Innovative SMEs collaborating with others as % of SMEs	1/3	1/3	1/9	9/24	4.17%
	Weight of non-CIS indicators in Outputs	**Weight of indicator in non-CIS indicators**	**Weight in Outputs**	**Weight of Outputs in IUS**	**Weight of indicator in RIS**
3.2.1 Employment in knowledge-intensive services + Employment in medium-high/high-tech manufacturing as % of total workforce	4/7	100%	4/7	7/24	16.67%
	Weight of CIS indicators in Outputs	**Weight of indicator in CIS indicators**			
3.1.1 SMEs introducing product or process innovations as % of SMEs	3/7	33.33%	1/7	7/24	4.17%
3.1.2 SMEs introducing marketing or organisational innovations as % of SMEs	3/7	33.33%	1/7	7/24	4.17%
3.2.4 Sales of new to market and new to firm innovations as % of turnover	3/7	33.33%	1/7	7/24	4.17%

The combined weight of the CIS indicators would be 25%, identical to the weight of these indicators in the IUS. But the table also shows that some indicators have a weight 4 times that of the CIS indicators and this clearly overemphasizes the relative importance of these indicators.

The weights shown in Table B.1 have therefore been combined with a scheme of equal weights where each of the 12 indicators would receive a weight of 8.33%. The combination of weights results in the percentage share of each of the indicators in the regional innovation index as shown in Table B.2.

Table B.2: Percentage contribution indicators to RII, degree of skewness and transformation for each of the RIS indicators

	"RIS 2009 weights"	**"Equal weights"**	**RIS 2012 weights**
ENABLERS			
1.1.2 Percentage population aged 25-64 having completed tertiary education	16.67%	8.33%	**12.5%**
1.3.1 R&D expenditure in the public sector as % of GDP	16.67%	8.33%	**12.5%**
FIRM ACTIVITIES			
2.1.1 R&D expenditure in the business sector as % of GDP	8.33%	8.33%	**8.33%**
2.1.2 Non-R&D innovation expenditures as % of turnover	4.17%	8.33%	**6.25%**
2.2.1 SMEs innovating in-house as % of SMEs	4.17%	8.33%	**6.25%**
2.2.2 Innovative SMEs collaborating with others as % of SMEs	4.17%	8.33%	**6.25%**
2.2.3 Public-private co-publications per million population	8.33%	8.33%	**8.33%**
2.3.1 PCT patents applications per billion GDP (in PPS€)	8.33%	8.33%	**8.33%**
OUTPUTS			
3.1.1 SMEs introducing product or process innovations as % of SMEs	4.17%	8.33%	**6.25%**
3.1.2 SMEs introducing marketing or organisational innovations as % of SMEs	4.17%	8.33%	**6.25%**
3.2.1 Employment in knowledge-intensive services + Employment in medium-high/high-tech manufacturing as % of total workforce	4.17%	8.33%	**12.5%**
3.2.4 Sales of new to market and new to firm innovations as % of turnover	16.67%	8.33%	**6.25%**

All data will be normalized using the same procedure as in the IUS, where the normalized value is equal to the difference between the real value and the lowest value across all regions divided by the difference between the highest and lowest value across all regions. These values are first transformed using a power root transformation if the data are not normally distributed.

Most of the indicators are fractional indicators with values between 0% and 100%. Some indicators are unbound indicators, where values are not limited to an upper threshold. These indicators can have skewed data distributions (where most regions show low performance levels and a few regions show exceptionally high performance levels). For all indicators data will be transformed using a square root transformation with power *N* if the degree of skewness of the raw data exceeds 0.5 such that the skewness of the transformed data is below 0.5 (none of the imputed data are included in this process):

$$\tilde{\tilde{X}}_r = \sqrt[N]{X_r}$$

The data will then be normalized using the min-max procedure where the transformed score is first subtracted with the minimum score over all regions in 2006, 2008 and 2010 and then divided by the difference between the maximum and minimum scores over all regions in 2006, 2008 and 2010:

$$\hat{X}_r = \frac{\tilde{\tilde{X}}_r - MIN(\forall_r \tilde{\tilde{X}}_r)}{MAX(\forall_r \tilde{\tilde{X}}_r) - MIN(\forall_r \tilde{\tilde{X}}_r)}$$

The maximum normalised score is thus equal to 1 and the minimum normalised score is equal to 0. These normalised scores are then used to calculate the different composite indicators, including the regional innovation index using the weighting scheme as shown in Table B.2.