Unclassified **DSTI/DOC(2008)2**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **03-Jun-2008**

English text only

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY

DSTI/DOC(2008)2
Unclassified

THE OECD REGPAT DATABASE: A PRESENTATION

STI WORKING PAPER 2008/2
Statistical Analysis of Science, Technology and Industry

Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia and Dominique Guellec

JT03246949

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

English text only

STI Working Paper Series

The Working Paper series of the OECD Directorate for Science, Technology and Industry is designed to make available to a wider readership selected studies prepared by staff in the Directorate or by outside consultants working on OECD projects. The papers included in the series cover a broad range of issues, of both a technical and policy-analytical nature, in the areas of work of the DSTI. The Working Papers are generally available only in their original language – English or French – with a summary in the other.

Comments on the papers are invited, and should be sent to the Directorate for Science, Technology and Industry, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

The opinions expressed in these papers are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

http://www.oecd.org/sti/working-papers

THE OECD REGPAT DATABASE: A PRESENTATION

ABSTRACT

The OECD REGPAT database presents patent data that have been linked to regions according to the addresses of the applicants and inventors. The data have been 'regionalised' at a very detailed level so that more than 2 000 regions are covered across OECD countries. REGPAT allows patent data to be used in connection with other regional data such as GDP or labour force statistics, and other patent-based information such as citations, technical fields and patent holder's characteristics (industry, university, etc.), thus providing researchers with the means to develop a rich set of new indicators and undertake a broad range of analyses to address issues relating to the regional dimension of innovation. By making regionalised patent data available to all students interested in the field, the OECD aims to stimulate research and contribute to a better understanding of the regional dimension of innovation. In addition, the methodology used for the construction of REGPAT is published, to give users the opportunity to suggest modifications and thus contribute to improvements in the quality of REGPAT. The full technical description of the REGPAT database as accessible to users is provided in annex. Patent data provide unique insights into the outcome and characteristics of inventive activities, including at regional level. They have limitations however, like all data sources, and should be handled with methodological care.

BASE DE DONNÉES REGPAT DE L'OCDE
PRÉSENTATION

RESUMÉ

La base REGPAT de l'OCDE présente des données relatives aux brevets appariées à des régions en fonction des adresses des demandeurs et inventeurs. Le niveau de détail de cette « régionalisation » est très poussé, de sorte que plus de 2 000 régions de toute la zone OCDE sont couvertes. REGPAT permet d'utiliser les données concernant les brevets en relation avec d'autres données régionales telles que le PIB ou les statistiques sur la main-d'œuvre, et avec d'autres informations propres aux brevets – citations, domaines techniques, caractéristiques du détenteur du brevet (secteur d'activité, université, etc.) ; les chercheurs peuvent ainsi agencer à leur guise un ensemble élargi d'indicateurs nouveaux et se livrer à des analyses très diverses portant sur les questions liées à la dimension régionale de l'innovation. En mettant des données régionalisées sur les brevets à la disposition de tous les analystes qui s'intéressent à ce domaine, l'OCDE a pour objectif de stimuler la recherche et de concourir à mieux faire appréhender cette dimension. Par ailleurs, la méthodologie présidant à la construction de REGPAT est rendue publique, de sorte que ses utilisateurs peuvent suggérer des modifications et, par là, contribuer à son amélioration qualitative. La description technique complète de la base telle qu'y accède l'usager est fournie en annexe. Les données relatives aux brevets livrent des enseignements sans équivalents sur les résultats et les caractéristiques des activités d'invention, y compris au niveau régional. Comme toutes les sources de données, elles comportent toutefois des limites et doivent être manipulées avec les précautions méthodologiques d'usage.

TABLE OF CONTENTS

1. Background[1]

The OECD REGPAT database presents patent data that have been linked to regions according to the addresses of the applicants and inventors. The data have been ‘regionalised’ at a very detailed level so that more than 2 000 regions are covered across OECD countries. REGPAT allows patent data to be used in connection with other regional data such as GDP or labour force statistics, and other patent-based information such as citations, technical fields and patent holder’s characteristics (industry, university, etc.), thus providing researchers with the means to develop a rich set of new indicators and undertake a broad range of analyses to address issues relating to the regional dimension of innovation.

1.1. The regional dimension of innovation

Empirical evidence shows that innovative activities are not distributed evenly within countries, some regions being highly innovative while others do very little innovation. The development of innovation can be influenced by local/regional characteristics such as governance, infrastructure and factor endowment (for example, the availability of skilled labour) as much as by national policies such as those concerning research and development, protection of intellectual property rights (IPR) or competition.

Various types of “economies of agglomeration”, related notably to the use of common resources by innovative actors, can explain regional patterns of innovation. For instance, clusters may benefit from positive “local externalities” due to geographical proximity (lower communication and transportation costs). The local availability of skilled labour can also play a role, as it makes the skilled labour market more fluid. The presence of many firms developing similar technologies for similar markets can stimulate both co-operation and competition, both factors of innovation. The presence of high quality publicly funded research (universities, public laboratories) can generate knowledge spillovers and feed-in to the skilled labour market. Conversely, certain costs are associated with concentration, such as congestion. Hence, there are both advantages and drawbacks in the agglomeration of innovative activities.

This raises a number of policy-relevant questions:

- What are the characteristics which make certain regions perform better than others in innovative activities?
- What is the effect of innovation on local/regional economic performance – *i.e.* what are the linkages between the innovative sector and other sectors in a given locality or region?
- Is local/regional development favoured by specialisation (a source of economies of scale) or by diversity (cross-technical fields fertilisation)?
- What are the nature of linkages between successful innovative regions and neighbouring regions? Are they positive (knowledge spillovers) or negative (depletion of a common pool of resources)?
- What local/regional policies have proved successful in promoting the development of innovation?

1. The regionalisation procedure used to compile REGPAT (reported in section 2 of this document) was developed by Stéphane Maraut. Hélène Dernis (OECD) was responsible for the compilation and analysis of the indicators of sections 2 and 3. Vincenzo Spiezia, Colin Webb and Dominique Guellec (OECD) have contributed to REGPAT and to this document.

- Many governments have policies which aim to encourage "poles of excellence", but this may lead to a widening of disparities between regions. Should government encourage the geographical concentration of innovative activities or attempt instead to disperse them?

Describing and understanding regional patterns of innovation is important both for regional and national policy makers: for regional policy makers as it provides them with benchmarks and references; for national policy makers as it captures an important dimension of the national innovation policies. In other words, issues like the mobility of human capital, networks of excellence or university-industry linkages, which are major themes of national innovation policies, have a spatial dimension which should not be ignored.

1.2. The OECD Regional Database

The OECD Regional Database (RDB) provides quantitative information on socio-economic issues in 2 014 regions within 30 OECD member countries. The database includes regional statistics on four major topics (demographics, regional accounts, labour market, and social issues) and a derived indicator on the typology of regions. The database contains annual data from 1990 to 2006, although the latest available year varies depending on variables and countries.

The RDB has been established to provide an internationally comparable database for the analysis of economic, institutional and environmental issues at the sub-national level. In any analytical study conducted at sub-national levels, the choice of the territorial unit is of prime importance. To address this issue, regions within each member country have been classified in two territorial levels (TLs). The higher, more aggregate, Territorial Level 2 (TL2) consists of about 335 macro-regions while the lower, more detailed Territorial Level 3 (TL3) is composed of 1 679 micro-regions.

This classification, which for European countries is largely consistent with the Eurostat classification, allows greater comparability of regions at the same territorial level. Indeed, these two levels, which are officially established and relatively stable in all member countries, are used by many as a framework for implementing regional policies. The differences with the Eurostat classification of regions, NUTS[2], concern Belgium, Greece and the Netherlands where the NUTS 2 level corresponds to the OECD TL3. For the United Kingdom the Eurostat NUTS1 corresponds to the OECD TL2. For Denmark, where the NUTS 2 has not been established, the OECD TL2 corresponds to groups of TL3/NUTS3 regions.

Due to limited data availability, labour market indicators in Canada and Australia are presented for a different grid (groups of TL3 regions in the case of Canada). Since these breakdowns are not part of the OECD official territorial grids, for the sake of simplicity they are labelled as Non Official Grids (NOGs).

Table 1 summarises the OECD territorial grids. For each country, data are collected at two sub-national levels + NOGs:

- Territorial level 2 (TL2), which refers to the 335 large regions of the OECD area.
- Territorial Level 3 (TL3), which refers to the 1 679 small regions of the OECD area.
- Non Official Grids, Australia (30) and Canada (71).

2. The Nomenclature of Territorial Units for Statistics (NUTS) was established by Eurostat more than 30 years ago in order to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union. Further details at: http://ec.europa.eu/eurostat/ramon/nuts/splash_regions.html

Table 1. Territorial grids by country

Country	Large Regions (TL2)	Small Regions (TL3)	Non-Official Grids (NOGs)
Australia	8 States/Territories	58 Statistical Divisions	30 LFS Dissemination Regions
Austria	9 Bundesländer	35 Gruppen von Politischen Bezirken	-
Belgium	3 Régions	11 Provinces	-
Canada	12 Provinces and Territories	288 Census Divisions	71 LFS Economic Areas
Czech Republic	8 Groups of Kraje	14 Kraje	-
Denmark	3 Regions	15 Amter	-
Finland	5 Suuralueet	20 Maakunnat	-
France (without DOM-TOM)	22 Régions	96 Départements	-
Germany	16 Länder	97 Spatial planning regions (groups of Kreise)	-
Greece	4 Groups of Development regions	13 Development regions	-
Hungary	7 Tervezesi-statisztikai regio	20 Megyek (+Budapest)	-
Iceland	2 regions	8 Landsvaedi	-
Ireland	2 Groups Regional Authority Regions	8 Regional Authority Regions	-
Italy	21 Regioni	103 Province	-
Japan	10 Groups of prefectures	47 Prefectures	-
Korea	7 Regions	16 Special city, Metropolitan area and Province	-
Luxembourg	1 State	1 State	-
Mexico	32 Estados	209 Grupos de Municipios	-
Netherlands	4 Landsdelen	12 Provinces	-
New Zealand	2 Groups of regional Councils	14 Regional Councils	-
Norway	7 Landsdeler	19 Fylker	-
Poland	16 Voïvodships	45 Subregions	-
Portugal	5 Comissaoes de coordenaçao regional + 2 Regioes autonomas	30 Grupos de Concelhos	-
Slovak Republic	4 Zoskupenia Karajov	8 Kraj	-
Spain	19 Comunidades autonomas	52 Provincias	-
Sweden	8 Riksomraden	21 Län	-
Switzerland	7 Grandes régions	26 Cantons	-
Turkey	26 Regions	81 Provinces	-
United Kingdom	12 Government Office Regions + Countries	133 Upper tier authorities or groups of lower tier authorities or groups of unitary authorities or LECs or groups of districts	-
United States	51 States	179 BEA Economic Areas	-

The RDB collects regional statistics on geography, population, GDP, labour force, employment, education, health, environment and other social indicators. The database contains only one measure of regional innovation: the number of patent applications to the main patenting office of the country by the residence of inventors. This measure, therefore, covers only a share – although the largest one – of the total number of applications in a country. In addition, data are available only at a fairly aggregated level, *i.e.* TL2 (compiled from Patent offices' annual reports). Finally, the information is not available by the region of the residence of the applicants.

R&D expenditures are not collected because their availability at the regional level is very limited. A major reason for this is the difficulty to allocate R&D carried out by multi-establishment companies when establishments are located in different regions.

Other economic variables related to the analysis of innovation and collected in the RDB include: total GDP, GDP per worker and the tertiary-level attainment rate.

Total GDP is collected for small regions (TL3) but virtually no industry breakdown is available for either TL2 or TL3 regions. As for R&D, this is due to the difficulty to allocate value added produced by multi-establishment companies when establishments are located in different regions. In addition, the combination of industry breakdown and geographic delimitation may quickly lead to the identification of companies. Confidentiality issues, therefore, prevent the publication of detailed regional data.

GDP per worker is available for both TL2 and TL3 regions. It is defined as the ratio between the level of GDP produced in a region and the total number of workers employed in the same region. The latter, therefore, includes individuals working in the region but living elsewhere (in-commuters) and excludes individuals living in the region but working elsewhere (out-commuters). As such, GDP per worker is a correct measure of the average labour productivity in a region.

The tertiary-level attainment rate is defined as the number of persons in the 25-64 age group who have completed tertiary educational programmes as a percentage of all persons of the same age. Tertiary education includes both university studies and advanced professional programmes. These correspond to levels 5A, 5B and 6 of the International Standard Classification for Education (ISCED 97). This indicator is available for both TL2 and TL3 regions

1.3. Regional patent data

Patents are a means of protecting legally inventions developed by firms, institutions or individuals, and as such they may be interpreted as indicators of invention (see *OECD Patent Manual*, forthcoming, for a detailed account of the statistical properties of patents, the way to use patent indicators and the limits of this data). Before an invention can become an innovation, further entrepreneurial efforts are required to develop, manufacture and market it. Patents are aimed at ensuring market exclusivity to their holder on the protected invention: No third party is allowed to make any use of the invention without the consent of the patent holder. The patent is applied to a national patent office, for the corresponding national market. If a patent is granted (it can be refused), it is valid for a maximum of 20 years, after which the invention belongs in the public domain.

Patent indicators convey information on the *output* and *processes* of inventive activities. Patents protect inventions and, although the relationship is not simple, many studies have shown that after applying the proper controls there is a positive relationship between patent counts and other indicators related to inventive performance (productivity, market share, etc.). The relationship is not straightforward, it can vary across countries, industries and over time, but it can definitely be identified. Due to the richness of information reported in patent documents, statistical exploitation of the data can provide a unique insight into invention processes. Published patent documents reveal information on the technological content of the invention (notably its particular technical fields) and on the geographical location of the inventive process (via the addresses of the inventors and of the owners). By identifying the inventors and owners of inventions, patent data, when matched with complementary data, can provide insights into the organisation of the underlying research process (for example, alliances between firms, co-operation between firms and public research organisations, the respective role of multinationals and small firms, size and composition of research teams, etc.). Patents give unique information on the technical fields of inventions (biotechnology, wind energy, etc.). Patents can also reflect the type of output of inventors or their mobility and networks; and patents allow tracking the diffusion of knowledge (the influence of particular inventions on other, subsequent inventions).

Patent indicators can be used in conjunction with R&D and innovation survey data. As compared with R&D, patents are a measure of output (inventions); as compared with innovation, patents are a measure of input (an invention can be put to market or not). Much R&D is unsuccessful and won't result in a patent; many inventions won't go to the market and won't result in an innovation. Using together data on the three stages of the process (which is not a linear one, of course) gives a unique insight into the dynamics of technology and innovation.

1.4. Indicators based on regionalised patents

When compiling or analysing indicators with regionalised patents, it is necessary to have some characteristics of patents and some rules in mind, so as to make the best use of the information and not misinterpret the indicators (a detailed account is available in the *OECD Patent Manual*, forthcoming).

- *Inventor v. owner region*: Patent data can be regionalised on the basis of the address of either the inventor or the holder. The inventor's address usually indicates where the invention was made (often a laboratory or a research establishment, or the place of residence of the inventor) while the owner's address indicates where the holder (usually a company, university or other type of entity) has its headquarters. These two concepts have obviously different economic interpretation, especially as many patents are filed by large companies having several establishments located in different regions and countries.
- *Fractional v. whole counting*: Patents usually have several inventors and can have several owners. When regionalising patents, a patent with, say, inventors in two regions can be either attributed wholly to the two regions, or shared (with a total of shares of 100%) between the two regions. As a significant proportion of patents have inventors from different regions it is important to specify what rule is used, and when one is better, to use it. For instance, when comparing the performance of regions it is recommended to use fractional accounting, which *i)* attributes to each region its actual contribution to the invention; *ii)* when summed over all regions gives a total of 100%. On the other hand, when compiling an indicator like "share of patents with co-inventors from another region", it is recommended to use whole counting both at the numerator and the denominator.
- *Priority year*: It is the year of first filing for a patent; it is the closest to the actual date of invention, and should therefore be used as the reference date when compiling patent indicators aimed at reflecting technological achievements. Other dates (national application, publication or grant) are dependent on administrative procedures and can be one to ten years after the invention and thus misleading when interpreting the data.

There are various notions of patents: Patents *per se* are patent applications which have been granted by a patent office. However, granting of a patent usually occurs two to ten years after application, making grant-based statistics inevitably outdated. It is therefore more common to count patent applications. Patent applications are filed to patent offices, which are usually national offices (the major ones being the USPTO, for the US, and the JPO, for Japan) but can be a regional one (such as the European Patent Office). In addition, there exists an international procedure, the Patent Co-operation Treaty (PCT) which allows inventors to file "pre-applications" to many offices world-wide at a relatively low cost. The PCT is managed by the World Intellectual Property Organisation (WIPO). The first version of the REGPAT database includes patent applications to the EPO, to the PCT and to the USPTO (although data for the latter is not as complete). This choice is justified *i)* by the quality of the basic data (clean and complete addresses are not easily accessible from most other patent offices); *ii)* the international character of these two procedures, which makes the resulting data more comparable across countries.

When using regionalised patent data, two particular issues need to be kept in mind. First, regarding inventors, one has to be careful not to go to a too detailed level in certain large urban areas. The inventor might live in a different area code than the laboratory s/he works in (it will be then in a neighbouring area). Several inventors of the same invention might live in different zones of the same (large) city while they work at the same place. Hence, for large urban areas, made up of several detailed regions, it can be better to work data at a more aggregated level (*e.g.* level 2 instead of level 3). That can apply, in Europe, to the Paris or London areas for instance.

Second, a patent application may be filed by an affiliate of a firm, or co-filed by the firm and one of its affiliates. The address of the affiliate will appear in these cases and may not reflect the location of the entity actually controlling the patent.

2. REGPAT methodology and database description

This section describes the methodology developed to identify regions on the basis of addresses of the patent's inventor(s) or applicant(s). The method basically consists of an iterative procedure that matches postal codes and/or town names, identified in the addresses, with regions using a set of lookup tables (such as a postal code – NUTS3 correspondence). Possible sources of improvement are discussed, based on the success assessment of the regional allocation.

2.1. Methodology

2.1.1. Method of determining the NUTS3 code

The NUTS3 code for an address is determined by identifying at least one of two components: the postcode and/or the town. When the postcode is recognised in an address, the corresponding NUTS3 code(s) are looked for in the long and/or short postcodes table.[3] Priority is given to the correspondence with the long postcode since they are much more precise. When the procedure results in one NUTS3 code, and one only (*i.e.* with no breakdown), then the NUTS3 code is selected and the town is then identified for information purposes.[4] If, on the other hand, the identified post code relates to several NUTS3 codes (breakdown), then a search is carried out using the town name. If the town allows one or more NUTS3 codes to be identified and one of these NUTS3 codes is the same as one of the NUTS3 codes of the breakdown (the postcode always takes priority over the town), then this NUTS3 code is selected, failing which the NUTS3 code is assigned on the basis of the breakdown derived from the postcode.

If no postcode is identified in the address, a search is carried out on the basis of the town. If this produces a match with one, and only one, NUTS3 code, then this NUTS3 code is allocated otherwise no NUTS3 code is allocated.[5]

---

3. Some countries have references with only long postcodes, others with short postcodes and others with both.

4. While giving priority to the postcode rather than the town is not compulsory, it has proven more judicious and effective for two reasons:
   - Unlike a town name, a postcode is unique, which avoids uncertainty.
   - It is much easier to identify a postcode in an address than the name of a town: it complies with a clearly defined pattern (a series of 5 digits, for example).

   The postcode and the town can sometimes contradict each other regarding the NUTS3 code, in which case the NUTS3 code is derived from the postcode.

5. Unlike postcodes, there is no breakdown of NUTS3 codes on the basis of town names.

- *Multiple allocation of NUTS3 codes*:

  Several NUTS3 codes may be allocated to a short postcode or, less frequently (in the case of Canada and Australia), to a long postcode. This concept was identified in the source of data that was used to derive the postcode lookup tables and does not apply to town names. A town's NUTS3 code is identified when there is only one occurrence of the town's name in the country. Otherwise, it is difficult to determine with certainty which is the correct NUTS3 code for that town. As an example, several French towns named "Châtillon" are located in different regions (with different NUTS3 codes), which makes it impossible to properly identify the NUTS3 code from the name of the town in the address only.

- *Postcode in the address*:

  The identification of the post code in an address relies on the identification of one of the country's postcode patterns. In the case of France, for example, there is only one pattern – a series of 5 digits – while in the case of Romania three patterns coexist made up of 6, 5 or 4 digits. When a pattern is recognised, the long postcode is identified – the short postcode is deduced if required.

  When several postcodes are apparent in an address, the postcode the farthest on the right in the address is taken into account.

- *Name of the town in the address*:

  To identify the town in an address, a syntactic analysis of the address is performed on the basis of a reference table containing the town names for each country. For each address, from 0 to n towns can be found, *i.e.* when the street name is the name of a town (case 1), when a sub-component of the town name is also the name of a town (case 2), or when town names overlap (case 3). In every case, the town selected is the one ending farthest to the right. As an example, the name of the town is identified as follows:

Case 1: The town selected is the one "farthest on the right".

Address:	12 rue de Versailles Paris
Towns:	« Versailles » and « Paris »
Town selected:	Paris

Case 2: The town selected is the one that "contains the other", *i.e.* the biggest.

Address:	12 rue des Lilas Champagne au Mont d'Or
Towns:	« Champagne » and « Champagne au Mont d'Or »
Town selected:	Champagne au Mont d'Or

Case 3: The town selected is the one "farthest on the right".

Address:	12 rue des Lilas Saint Julien du Mont
Towns:	« Saint Julien » and « Julien du Mont »
Town selected:	Julien du Mont

2.1.2. Procedures for improving data

A number of procedures have been put in place to improve results: manual or automatic. Manual procedures are more time consuming despite being conceptually simpler and producing very good results. The first four procedures below are very simple and with a low risk of errors, whereas the last one (deduction) is based on assumptions, with higher risks of error.

Automatic procedures are more complex to implement, but can easily be replicated and have a higher productivity level. However, the resulting gains are small or negligible. Furthermore, all these procedures are based on initial assumptions involving a risk of error (which can be preset).

2.1.2.1. Manual procedures

- *Cleaning up special characters and abbreviations in addresses and town names*:

  In this context, addresses and town names are treated in the same way. Cleaning up of special characters is performed by replacing special characters with the generic characters (*e.g.*: é→ e, à →a). Some letter combinations are also replaced by generic words (*e.g.* St. = Saint).

  Example:

  St-Etienne becomes Saint Etienne.

  Note: Parentheses () are treated as a special case.

- *Specific treatment of names of reference towns*:

  Reference data on town names include special cases that need to be properly addressed. In certain Spanish regions for example, different languages in which the town's name is spelled are separated with "/". The town name is consequently divided into several town names.

  Example:

  Alicante/Alacant gives 2 towns: Alicante and Alacant, with the same NUTS3 code.

- *Languages*:

  Languages in which addresses are written are the language of the country (English and French in the majority of cases). However, some reference town tables were provided in the language of the country only. The reference tables were complemented with basic translations of the town's names for towns having the most frequent occurrences.

  Examples:

  München is also named Munich

  Wien is also named Vienne or Vienna

- *Addition of postcode patterns*:

  Even though the postcode patterns are property identified for a country, writing habits of the postal codes can lead to slightly different patterns depending on the country. Pattern variations are often due to addition of a space or a dash character within the postcode. Specific patterns were therefore added for some countries.

  Example:

  75011 can be written 75 011; 98789 can be written 98-789.

- *Deductions*:

  When several occurrences of a same town name are found in the lookup table with different NUTS codes, a combination of town name/NUTS3 code can be automatically allocated as the correct reference by deactivating the other possible combinations. This procedure has been used in the case of US addresses: when the address gives "Washington" only, it is assumed that the reference town is Washington DC.

2.1.2.2. Automatic procedures

- *Learning by doing*:

  Certain towns found in addresses are not referenced among the reference towns but they may frequently occur in addresses field. For such occurrences, NUTS codes may be identified using the post codes. Therefore, if the NUTS codes are identical for a significant number of addresses, then the combination of town / NUTS code is added in the lookup table of town names.

- *Name recognition*:

  A degree of recognition of a town name can be determined in an unresolved address by comparing the whole address with the reference town names. However, such data processing is costly and only minor improvements were realised when tested on a small sample. Therefore, this procedure was not implemented on the whole set of unresolved addresses.

  Example:

  In the address "12, rue des trois Fontanot Nanterre", the town of Nanterre is recognised at more than 95% of occurrences and is eligible.

- *Linking inventors and applicants addresses*:

  This linking process consists of examining whether an unidentified address of inventor (applicant) could be compared with data on other patent publications originating from the same inventor (applicant), in which a region code was identified. If so, it is likely that the region codes are the same. However, such cases are rare among the unidentified addresses and the improvements made by this procedure are negligible.

- *Merging results from different sources*

  Findings based on different patent data sources could be combined to improve the success rate of regionalisation. However, data sources are very similar and improvements in the results are negligible.

2.2. Data sources

2.2.1. Addresses

The regionalisation procedures described above were conducted on the addresses extracted from the following data sources:

- EPO's Worldwide Statistical Patent database (PATSTAT), April and October 2007 editions: extraction of patents taken at the EPO, the USPTO and PCT filings (WO publications).
- Inventors and Applicants records from EPO patents (data extracted from Epoline web services).

Regionalisation process was conducted on a list of 36 countries, mainly OECD member countries and selected non-member countries from Europe.

2.2.2. NUTS3 code

Reference tables matching towns and long/short postcodes with NUTS codes were developed using data provided by Eurostat (for European countries) and OECD's Directorate on Public Governance and Territorial Development (GOV).

2.2.3. Data availability by country

For a given country, the identification of the regional code can be derived from three complementary lookup tables based on: the town names, the long postcode and the short postcode (see Table 2 below). Note that small countries such as Luxembourg and Cyprus[6] are treated as one single region for this analysis.

6. ***Footnote by Turkey:*** The information in this document with reference to "Cyprus" relates to the southern part of the island. There is no single authority representing both Turkish and Greek Cypriot people on the island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

   Footnote by all the European Union Member States of the OECD and the European Commission: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Table 2. Complementary sources of data by country

	Country code	Town	Long postcode	Short postcode
Australia	AU	✓	✓	-
Austria	AT	✓	✓	✓
Belgium	BE	✓	✓	✓
Bulgaria	BG	✓	✓	✓
Canada	CA	✓	✓	✓
Croatia	HR	✓	✓	✓
Cyprus	CY	✓	-	-
Czech Republic	CZ	✓	✓	✓
Denmark	DK	✓	✓	✓
Estonia	EE	✓	✓	✓
Finland	FI	✓	✓	✓
France	FR	✓	-	✓
Germany	DE	✓	✓	✓
Greece	GR	✓	✓	✓
Hungary	HU	✓	✓	✓
Ireland	IE	✓	-	-
Italy	IT	✓	✓	✓
Japan	JP	✓	✓	-
Korea	KR	✓	✓	✓
Latvia	LV	✓	✓	✓
Lithuania	LT	✓	✓	✓
Luxembourg	LU	✓	-	-
Malta	MT	✓	✓	-
Netherlands	NL	✓	✓	✓
Norway	NO	✓	✓	✓
Poland	PL	✓	-	✓
Portugal	PT	✓	✓	✓
Romania	RO	-	-	✓
Slovak Republic	SK	✓	✓	✓
Slovenia	SI	✓	-	✓
Spain	ES	✓	-	✓
Sweden	SE	✓	✓	✓
Switzerland	CH	✓	✓	✓
Turkey	TR	-	✓	✓
United Kingdom	GB	✓	✓	-
United States	US	✓	✓	-

2.3. Results[7]

The success of the regionalisation procedure highly depends on the patent office and on the country analysed. Addresses provided in EPO patents are more complete than those of USPTO and of PCT (WO): in most cases, both the town name and the postal codes are available in the address field of EPO patents. In USPTO patents, the postal codes are often missing and the regionalisation process is mostly based on the town's names. In PCT data extracted from PATSTAT, the addresses field is almost not available (according to Table 5, almost 90% of PCT addresses are missing).

Based on EPO data (see Tables 3, 6 & 7), the regionalisation process reached an average success rate of 98% (average based on the 36 countries covered). The success rate is smaller for countries such as Croatia, Malta, Portugal, Romania and Turkey. Among the large OECD patenting countries, Canada and the United Kingdom are the two countries with a lower rate of regionalised data (between 90% and 95%). For these countries, further work should be undertaken to improve the lookup tables, especially those based on towns' names.

Regarding the regionalisation process itself, the most frequently applied method is based on postal codes patterns. Regionalisation is mostly based on towns' names for countries such as Ireland, Japan and Korea, as well as for countries from Eastern Europe. In some European countries, the regionalisation process ended up in allocations of NUTS codes with a breakdown. The share of addresses with a NUTS3 breakdown is 14% for Germany in PATSTAT (see Table 3). This issue needs to be addressed in order to get a unique correspondence between an address and a single NUTS code.

In the case of USPTO data, the overall success rate stands at 72% of all valid addresses (see Table 4). Most of the regionalisation process is based on the towns' names comparisons, due to the structure of the addresses itself (almost no postal codes recorded in the address field). Addresses, as other textual information, are more subject to spelling mistakes than other fields. In order to enhance the success rate, further work would need to be conducted in either cleaning the addresses or integrating more name variations in the lookup tables.

The number of patents that have no region allocated is evolving over time, notably at the EPO. As shown in Table 8, the share of patents with no region allocated is decreasing in later years, notably in addresses from Canada, Poland, Portugal and Turkey where the shares lost more than 10 percentage points between 1985 and 2004. However, such a trend is not observed in USPTO records: only a few countries' addresses see notable improvements in the regional allocation over time, such as Australia, Ireland and the United States.

Overall, among all data sources, several sources of failure of the regional allocation procedure have been identified:

- In most cases, the address provides a town name that exists in different regions.
- The address refers to a wrong country.
- The address refers to a postcode and/or a town which are not listed in the reference data.
- The address field is empty or not valid.

7. This section reports the results of the regionalisation exercise as conducted in the initial phase of REGPAT, early 2008. Following versions of REGPAT will include improved rates of regionalisation and an extended coverage of countries.

Table 3. Number of inventors and applicants' addresses in EPO patents (PATSTAT)

	Total	*of which* valid addresses	Successful regionalisation (%)	*% of which with*			
				Zip code	Town name	Breakdown	Mixed procedure
Australia	28 672	28 273	97.9	91.2	5.3	1.4	0.0
Austria	34 821	30 282	97.8	86.2	11.6		
Belgium	43 355	37 635	99.1	94.7	1.9	2.4	0.0
Bulgaria	1 183	1 138	***92.7***	40.2	51.8	0.8	
Canada	51 616	51 080	***94.9***	81.1	10.1	3.3	0.4
Croatia	668	609	***68.3***	68.0		0.3	
Cyprus	254	228	100.0				
Czech Republic	2 156	1 961	97.9	89.6	4.7	3.3	0.2
Denmark	25 461	23 740	99.5	82.0	10.1	7.4	
Estonia	270	234	96.2	47.0	49.1		
Finland	29 516	28 367	99.7	98.6	0.9	0.3	
France	247 702	236 018	99.7	98.9	0.7	0.1	0.0
Germany	682 224	575 299	99.6	82.4	2.6	13.7	0.9
Greece	1 895	1 728	98.4	87.8	9.3	1.3	
Hungary	10 244	10 000	99.7	93.5	5.4	0.8	
Ireland	7 146	6 531	95.2		95.2		
Italy	125 029	110 818	99.6	91.6	7.0	0.9	
Japan	833 426	831 761	99.4	15.1	84.3	0.0	
Korea	57 844	57 732	99.1	45.4	53.7		
Latvia	235	217	99.1	91.2	5.5	2.3	
Lithuania	194	178	99.4	39.9	23.6	36.0	
Luxembourg	3 053	2 567	100.0				
Malta	144	116	***81.9***		81.9		
Netherlands	86 873	79 095	99.1	94.5	1.6	3.0	0.0
Norway	11 155	10 787	99.5	93.6	1.5	4.4	
Poland	3 033	2 892	***78.0***		78.0		
Portugal	1 403	1 285	***62.3***	42.9	15.5	4.0	
Romania	452	415	***31.8***	31.1		0.7	
Slovak Republic	578	477	98.5	91.4	3.4	3.8	
Slovenia	1 592	1 292	***94.6***	32.5	14.9	47.1	
Spain	26 759	24 309	99.5	92.0	7.6		
Sweden	59 361	56 261	99.5	92.3	0.6	6.6	0.0
Switzerland	80 487	70 914	99.5	58.0	36.4	4.7	0.3
Turkey	1 501	1 316	***74.7***	46.1		28.6	
United Kingdom	211 610	204 119	***93.4***	78.5	14.9		
United States	996 236	985 973	98.9	94.9	4.0		
Total	3 797 988	3 580 100	98.8	68.4	24.6	2.6	0.2

Sources: REGPAT and PATSTAT (October 2007) databases, February 2008.

Table 4. Number of inventors and applicants' addresses in USPTO patents (PATSTAT)

	Total	*of which* valid addresses	Successful regionalisation (%)	*% of which with*			
				Zip code	Town name	Breakdown	Mixed procedure
Australia	42 451	39 964	***65.8***	15.1	50.3	0.4	0.0
Austria	20 093	19 286	***83.4***	12.5	69.0	1.8	
Belgium	24 677	23 523	***78.4***	8.2	69.8	0.4	0.0
Bulgaria	1 758	1 737	***86.6***	5.0	81.6		
Canada	146 223	140 145	***69.5***	3.8	65.3	0.3	0.0
Croatia	456	434	***7.8***	7.8			
Cyprus	196	174	100.0				
Czech Republic	3 576	3 474	***77.5***	5.3	72.0	0.3	
Denmark	18 848	17 827	***63.0***	9.6	51.6	1.8	
Estonia	198	184	***89.1***	3.8	85.3		
Finland	23 226	22 341	***88.4***	10.6	77.7	0.1	
France	158 859	152 464	***83.2***	6.7	74.0	2.3	0.2
Germany	379 882	365 374	***78.3***	7.0	68.5	2.7	0.1
Greece	1 260	1 222	***81.1***	14.7	66.2	0.2	
Hungary	8 470	8 344	***94.9***	10.8	84.1	0.0	
Ireland	7 139	6 874	***61.5***		61.5		
Italy	65 888	62 300	***91.7***	9.0	82.5	0.2	
Japan	899 225	870 433	***91.1***	0.5	90.6		
Korea	127 227	123 816	***90.4***	1.9	88.5		
Latvia	140	135	***88.9***	14.8	74.1		
Lithuania	167	154	***87.7***	1.3	85.7	0.6	
Luxembourg	1 928	1 668	100.0				
Malta	74	71	***85.9***		85.9		
Netherlands	53 257	50 155	***76.4***	6.0	70.2	0.2	
Norway	11 025	10 526	***63.3***	16.3	45.8	1.2	
Poland	2 777	2 718	***79.9***		79.9		
Portugal	806	768	***34.9***	2.1	32.8		
Romania	597	587	***2.0***	2.0			
Slovak Republic	283	265	***81.9***	14.3	66.8	0.8	
Slovenia	717	686	***73.2***	7.4	61.4	4.4	
Spain	15 697	14 974	***92.8***	3.1	82.3	7.3	
Sweden	55 847	53 202	***73.7***	14.2	57.4	2.2	
Switzerland	55 533	52 505	***82.0***	7.8	70.3	3.8	0.1
Turkey	754	703	***5.4***	3.3		2.1	
United Kingdom	175 165	166 334	***51.8***	5.5	46.3		
United States	2 771 487	2 663 643	***61.1***	2.3	58.7		
Total	7 088 541	5 202 733	***72.4***	3.0	62.7	0.4	0.0

Sources: REGPAT and PATSTAT (October 2007) databases, February 2008.

Table 5. Number of inventors and applicants' addresses in PCT patents (PATSTAT)

	Total	*of which* valid addresses	Successful regionalisation (%)	*% of which with*			
				Zip code	Town name	Breakdown	Mixed procedure
Australia	38 326	5 427	98.4	94.9	1.2	2.2	0.0
Austria	12 820	1 613	99.8	82.9	0.4	16.4	
Belgium	13 561	1 656	98.4	90.3	2.9	5.2	
Bulgaria	680	25	***92.0***	44.0	48.0		
Canada	36 044	4 094	***90.9***	83.4	3.3	4.2	0.0
Croatia	704	31	***0.0***				
Cyprus	211	16	100.0				
Czech Republic	1 945	47	***87.2***	10.6	76.6		
Denmark	17 131	2 726	99.8	82.5	0.2	17.1	
Estonia	257	6	***33.3***		33.3		
Finland	19 056	2 718	99.9	99.0	0.5	0.4	
France	80 555	10 512	99.7	71.2	0.4	26.4	1.7
Germany	170 643	23 353	99.6	57.9	1.4	38.7	1.5
Greece	1 246	121	97.5	92.6	3.3	1.7	
Hungary	6 372	1 340	99.9	97.2	1.7	1.0	
Ireland	4 583	326	***48.2***		48.2		
Italy	31 469	2 893	99.6	94.8	3.6	1.2	
Japan	184 955	20 963	***63.7***		63.7		
Korea	35 952	545	***86.8***	81.5	5.3		
Latvia	261	5	***0.0***				
Lithuania	144	5	***80.0***	20.0		60.0	
Luxembourg	919	92	100.0				
Malta	58						
Netherlands	31 488	2 595	96.6	89.3	2.0	5.3	
Norway	10 858	1 489	99.7	90.9	0.3	8.4	
Poland	2 326	180	***73.9***		73.9		
Portugal	948	27	***44.4***		44.4		
Romania	590	34	***61.8***	61.8			
Slovak Republic	618	11	***90.9***	18.2	72.7		
Slovenia	1 146	1	***0.0***				
Spain	15 451	765	98.4	18.2	6.1	74.1	
Sweden	38 207	6 394	99.5	87.0	0.2	12.3	
Switzerland	26 257	2 932	99.5	66.2	0.4	32.7	0.2
Turkey	1 553	5	***20.0***			20.0	
United Kingdom	102 976	18 285	***90.9***	87.5	3.5		
United States	519 014	75 413	97.5	96.6	0.9		
Total	1 882 678	205 249	***94.0***	69.3	7.8	7.6	0.3

Sources: REGPAT and PATSTAT (October 2007) databases, February 2008.

Table 6. Number of inventors addresses in EPO data[1]

	Total	*of which* valid addresses	Successful regionalisation (%)	*% of which with*			
				Zip code	Town name	Breakdown	Mixed procedure
Australia	46 682	46 651	98.4	93.5	3.2	1.7	0.0
Austria	43 084	42 960	97.1	86.5	10.6		
Belgium	48 362	48 331	99.0	94.9	1.7	2.4	0.0
Bulgaria	845	845	***94.8***	72.5	21.4	0.8	
Canada	82 827	82 775	95.6	83.8	7.7	3.6	0.4
Croatia	899	896	***84.9***	84.6		0.3	
Cyprus	168	168	100.0				
Czech Republic	2 956	2 931	97.4	91.2	3.9	2.2	0.1
Denmark	32 851	32 814	99.5	81.8	9.0	8.8	
Estonia	323	323	***93.2***	43.3	49.8		
Finland	47 212	47 127	99.7	99.0	0.6	0.2	
France	302 475	302 281	99.5	98.8	0.6	0.1	0.0
Germany	940 797	940 200	99.0	84.8	2.6	10.9	0.7
Greece	2 061	2 057	97.7	88.0	8.6	1.1	
Hungary	12 719	12 697	99.2	95.4	3.3	0.5	
Ireland	8 021	8 003	***94.5***		94.5		
Italy	125 173	125 106	99.5	92.3	6.4	0.9	
Japan	682 844	682 258	97.9	20.0	77.9	0.0	
Korea	92 253	92 214	98.7	65.8	32.8		
Latvia	360	360	99.2	92.8	2.5	3.9	
Lithuania	309	309	97.4	41.7	23.6	32.0	
Luxembourg	2 923	2 922	100.0				
Malta	106	106	***79.2***		79.2		
Netherlands	95 286	95 205	98.9	96.1	1.5	1.3	0.0
Norway	15 691	15 533	99.3	93.4	1.4	4.5	
Poland	3 809	3 799	***72.6***		72.6		
Portugal	1 433	1 430	***64.6***	48.3	13.1	3.3	
Romania	598	598	***36.6***	36.3		0.3	
Slovak Republic	731	722	98.3	93.5	2.4	2.5	
Slovenia	1 939	1 936	96.5	28.1	8.1	60.3	
Spain	25 689	25 665	99.1	92.2	6.9		
Sweden	86 369	86 300	99.6	93.8	0.4	5.4	0.0
Switzerland	105 939	105 808	99.3	59.1	34.9	5.0	0.4
Turkey	1 626	1 624	***69.5***	45.0		24.5	
United Kingdom	237 390	237 266	***90.8***	78.6	12.1		
United States	1 661 608	1 659 970	99.0	96.1	3.0		
Total	4 888 846	4 868 271	98.4	77.1	15.3	2.6	0.2

1. Includes EPO patent applications (Direct EPO + Euro PCT at regional phase) and Euro PCT at international phase.

Sources: REGPAT and OECD patent databases, February 2008.

Table 7. Number of applicants addresses in EPO data[1]

	Total	*of which* valid addresses	Successful regionalisation (%)	*% of which with*			
				Zip code	Town name	Breakdown	Mixed procedure
Australia	27 610	27 608	98.0	90.9	5.7	1.4	0.0
Austria	23 774	23 772	98.0	77.4	20.7		
Belgium	21 726	21 726	99.6	97.6	0.8	1.3	
Bulgaria	523	522	96.2	75.1	19.7	1.3	
Canada	36 165	36 162	97.2	88.2	7.2	1.5	0.3
Croatia	590	590	***86.6***	85.1		1.5	
Cyprus	443	443	100.0				
Czech Republic	1 450	1 450	***94.7***	82.8	4.8	6.8	0.3
Denmark	17 928	17 928	99.9	86.1	8.4	5.3	
Estonia	120	120	98.3	62.5	35.8		
Finland	26 571	26 569	99.9	99.4	0.4	0.2	
France	170 989	170 986	99.4	99.0	0.3	0.1	0.0
Germany	436 295	436 283	99.4	95.1	1.7	2.4	0.2
Greece	1 353	1 353	97.5	89.9	6.6	1.0	
Hungary	3 820	3 818	99.6	94.9	3.5	1.3	
Ireland	4 889	4 888	98.2		98.2		
Italy	73 064	73 057	99.7	96.2	2.6	0.9	
Japan	436 056	436 015	99.6	17.6	82.0		
Korea	44 194	44 193	99.4	68.6	30.8		
Latvia	168	168	96.4	85.7	7.7	3.0	
Lithuania	92	92	95.7	37.0	41.3	17.4	
Luxembourg	3 207	3 207	100.0				
Malta	143	143	***81.8***		81.8		
Netherlands	81 535	81 534	99.5	94.8	0.4	4.3	
Norway	9 285	9 284	99.7	95.6	1.2	2.9	
Poland	1 682	1 682	***72.9***		72.9		
Portugal	804	804	***61.4***	39.7	15.0	6.7	
Romania	309	309	***41.4***	40.5		1.0	
Slovak Republic	366	366	99.5	86.9	4.6	7.9	
Slovenia	884	884	96.7	24.7	33.9	38.1	
Spain	14 442	14 439	99.5	94.1	5.4		
Sweden	54 148	54 147	99.8	95.3	0.8	3.8	0.0
Switzerland	80 399	80 396	99.3	64.1	31.6	3.6	0.1
Turkey	1 146	1 146	***78.7***	53.5		25.2	
United Kingdom	126 272	126 266	96.8	90.6	6.2		
United States	794 095	794 040	99.4	97.2	2.2		
Total	2 589 104	2 588 273	99.2	77.2	17.4	0.9	0.0

1. Includes EPO patent applications (Direct EPO + Euro PCT at regional phase) and Euro PCT at international phase.

Sources: REGPAT and OECD patent databases, February 2008.

Table 8. Share of patents with no regional allocation

by priority date and inventors' country

	EPO[1]					USPTO			
	1985	1990	1995	2000	2004	1985	1990	1995	2000
Australia	2.9	2.1	0.9	1.4	0.8	30.1	25.7	33.3	25.3
Austria	2.3	2.0	1.0	1.0	0.5	9.0	11.7	17.9	15.1
Belgium	3.6	2.0	0.7	0.3	0.2	22.8	14.7	22.1	21.1
Canada	14.1	9.7	5.6	4.4	3.7	33.7	32.9	34.6	32.4
Czech Republic	0.0	0.0	1.8	2.2	2.3	24.0	13.8	24.4	18.2
Denmark	3.4	0.8	0.4	0.4	0.4	33.8	41.7	50.7	44.6
Finland	2.2	0.3	0.7	0.2	0.5	14.1	9.5	14.1	10.1
France	0.9	1.1	0.5	0.2	0.3	15.9	15.2	18.9	15.5
Germany	2.5	2.7	1.0	0.2	0.1	17.8	19.3	21.6	21.2
Greece	0.0	1.3	0.0	2.3	0.0	30.9	27.1	9.3	21.5
Hungary	1.3	0.9	5.3	0.8	1.1	1.5	0.9	9.8	9.0
Ireland	5.8	5.8	3.9	6.0	6.0	43.3	43.0	37.4	33.4
Italy	1.3	0.8	0.3	0.2	0.3	7.0	7.3	10.3	5.9
Japan	1.3	1.0	0.7	0.3	4.8	8.0	6.7	8.7	5.1
Korea	5.6	3.5	1.9	1.5	3.3	7.4	8.8	11.3	6.9
Netherlands	2.1	1.3	1.0	0.3	0.5	20.0	22.4	24.5	20.5
Norway	2.4	1.6	2.7	1.1	0.9	35.0	30.3	40.4	36.4
Poland	36.4	26.8	19.9	20.2	20.8	29.1	8.4	17.1	16.4
Portugal	47.4	75.3	69.0	38.0	23.6	63.6	93.7	90.3	73.0
Slovak Republic			0.0	0.0	6.1			21.8	5.7
Spain	1.4	1.4	0.6	1.0	0.6	7.5	4.1	10.2	6.3
Sweden	0.6	0.1	0.4	0.4	0.5	28.9	32.7	37.4	29.4
Switzerland	1.4	1.3	0.7	0.6	0.3	13.8	16.0	17.2	15.8
Turkey	50.0	60.3	65.9	33.8	17.4	100	100	100	87.7
United Kingdom	21.4	9.1	4.3	3.4	3.2	48.4	50.7	51.9	48.0
United States	2.7	1.6	0.8	0.7	0.6	54.3	54.6	52.5	44.2

1. Direct EPO and Euro PCT at regional phase.

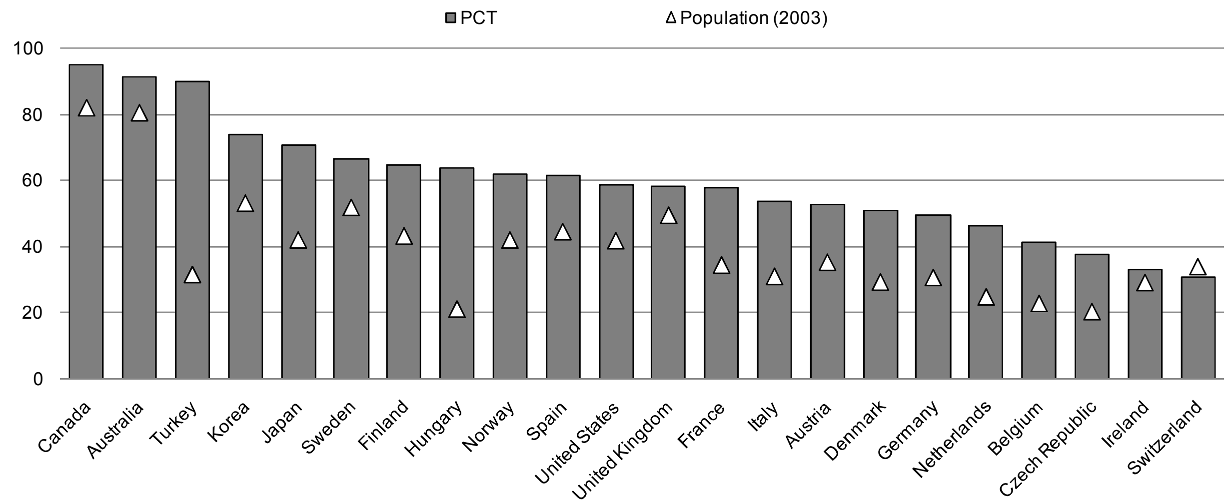
Sources: REGPAT and OECD patent databases, February 2008.

3. A sample of indicators from REGPAT

This section presents a sample set of indicators at regional level, using the REGPAT database. These indicators are mostly derived from data on EPO filings and Euro-PCT at international phase and presented according to the priority dates. All indicators were calculated according to the region of residence of the inventors, using fractional counts, with the exception of indicators on international co-operation. The regional breakdown was incorporated in the calculations when multiple NUTS codes were allocated to a single address. Some indicators were compiled using data derived from the OECD Regional database. For calculation and presentation purposes (maps), patent counts at NUTS3 level had to be converted to territorial levels (TL3 or TL2).

Inventive activities are concentrated in a small number of regions: the degree of concentration is much higher than that of population for most countries, as depicted in Figure 1 below. Switzerland has the lowest concentration ratio (31), almost at the same level than the concentration of population (34). Inventive activities are the most highly concentrated in certain regions for Canada (95), Australia (91) and Turkey (90).

Figure 1. Geographic concentration of patent applications filed under the PCT, 2005



Note: The Geographic concentration index for the variable y is defined as : $\left[\sum_{i=1}^{N} |y_i - a_i| \Big/ 2\right] \times 100$

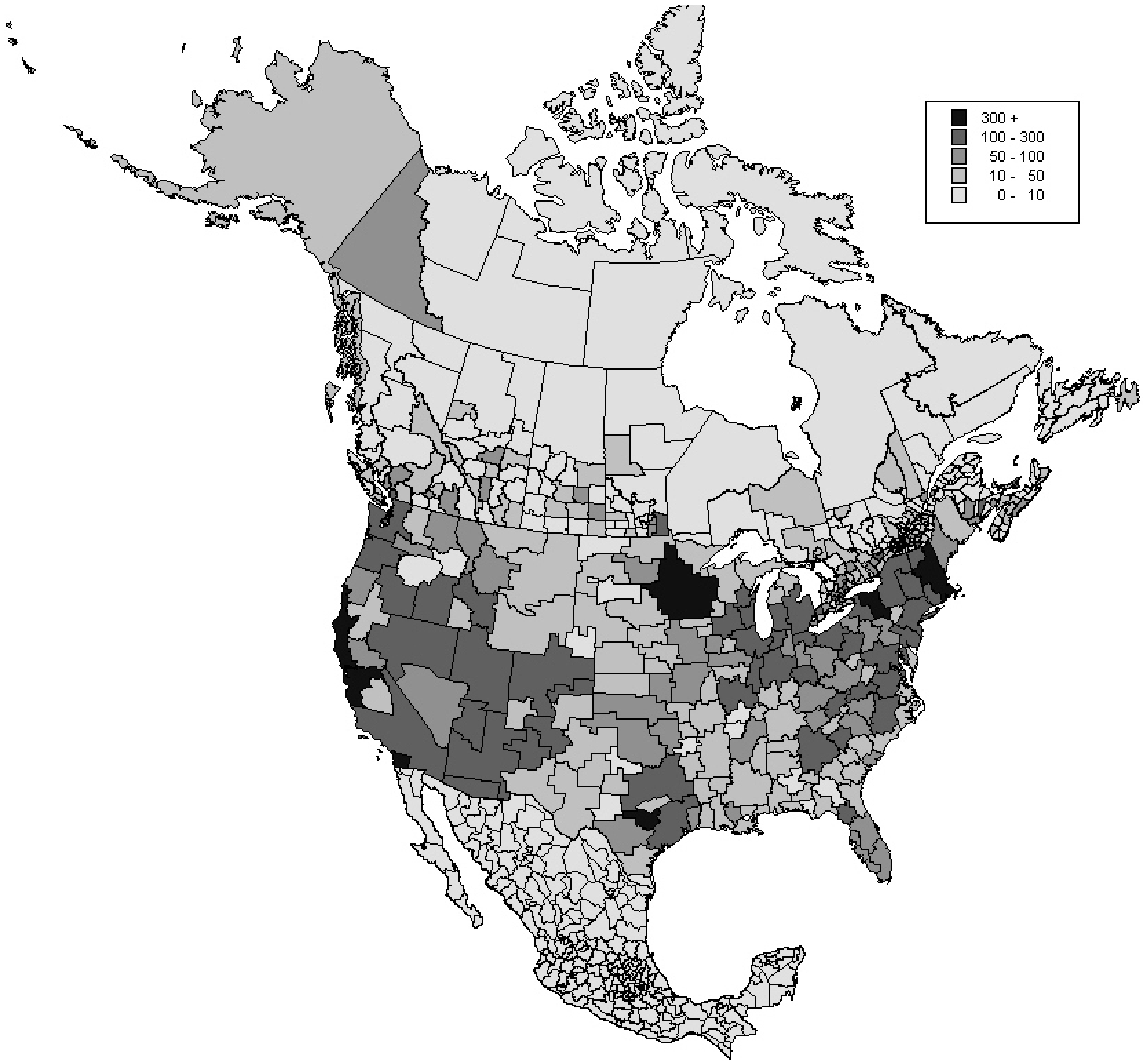
where y_i is the share of region i to the national total, a_i is the area of region I as a percentage of the country area, N stands for the number of regions and || indicates the absolute value. The index lies between 0 (no concentration) and 100 (maximum concentration) in all countries and is suitable for international comparisons of geographic concentration.[8] Only countries with more than 100 PCT applications in 2005 are included in the graph.

Sources: REGPAT, OECD Patent database and OECD Regional database, February 2008.

8. The value of the index of geographic concentration is affected by the size of regions. Therefore, differences in geographic concentration between countries may be partially due to differences in the average size of regions in each country.

Figure 2. Number of patent applications filed under the PCT per million population, 2004

North America

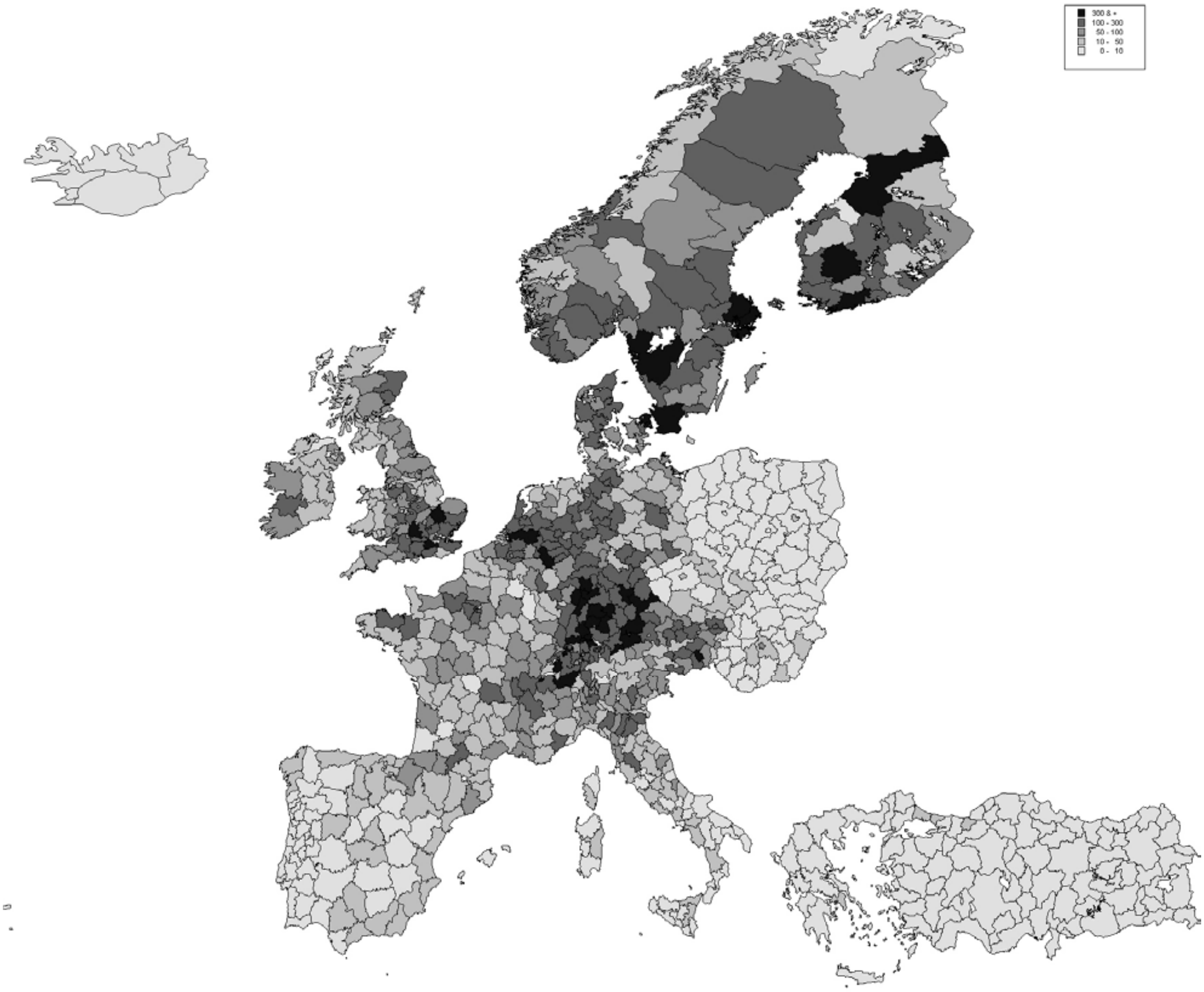


Sources: REGPAT, OECD Patent database and OECD Regional database, February 2008.

The regions of San Diego/Carlsbad/San Marcos and San Jose/San Francisco/Oakland were the most patent-intensive regions in 2004, as shown by the number of PCT filings per million population (with respectively 584 and 560 PCT filed per million inhabitants). The regions of Rochester/Batavia/Seneca Falls, Boston/Worcester/Manchester, Austin/Round Rock and Minneapolis follow with more than 300 PCT filings per million population.

Figure 3. Number of patent applications filed under the PCT per million population, 2004

Europe

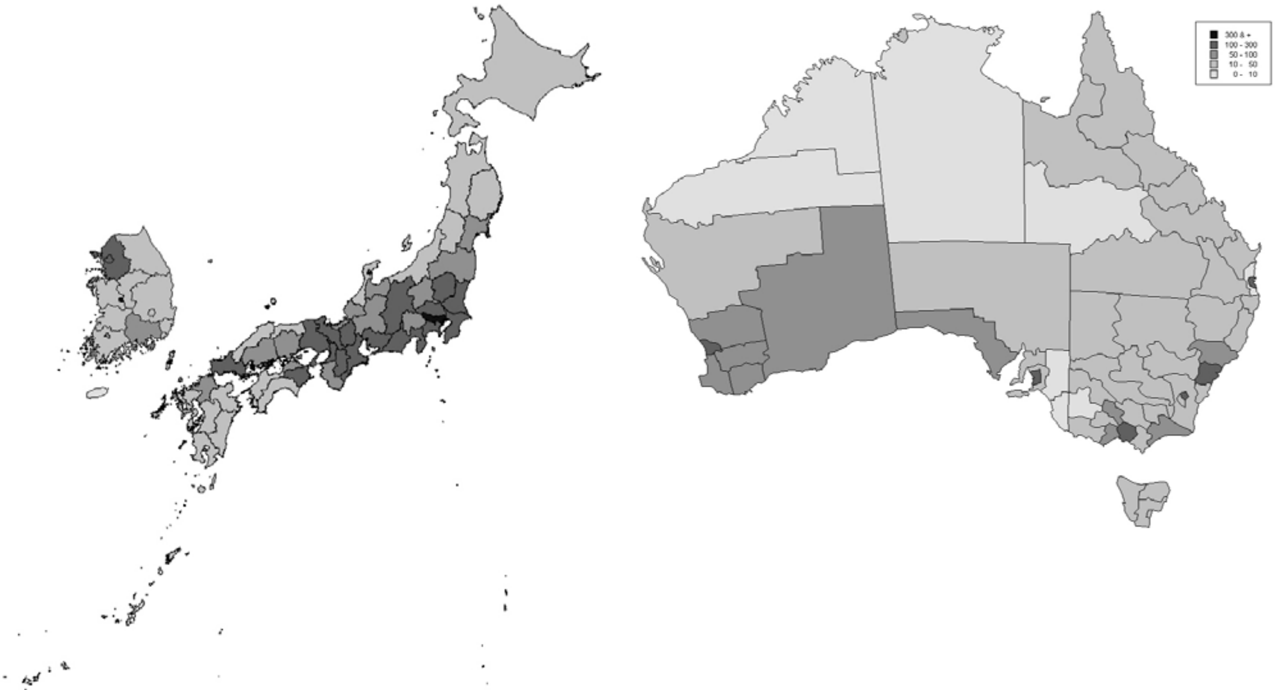


Sources: REGPAT, OECD Patent database and OECD Regional database, February 2008.

In Europe (Figure 3), the most patent-intensive regions are localised in the centre of Europe, in Nordic countries and in the United Kingdom. Provinces of Fribourg and Graubünden (Switzerland) lead with around 1 500 PCT filings per million inhabitants, followed by the regions of Noord-Brabant (766) in the Netherlands, Pirkanmaa (704) in Finland and Cambridgeshire (696) in the United Kingdom. The regions of Munich, Ostwürttemberg and Stuttgart in Germany are among the top innovative regions (over 500).

Figure 4. Number of patent applications filed under the PCT per million population, 2004

Asia-Pacific

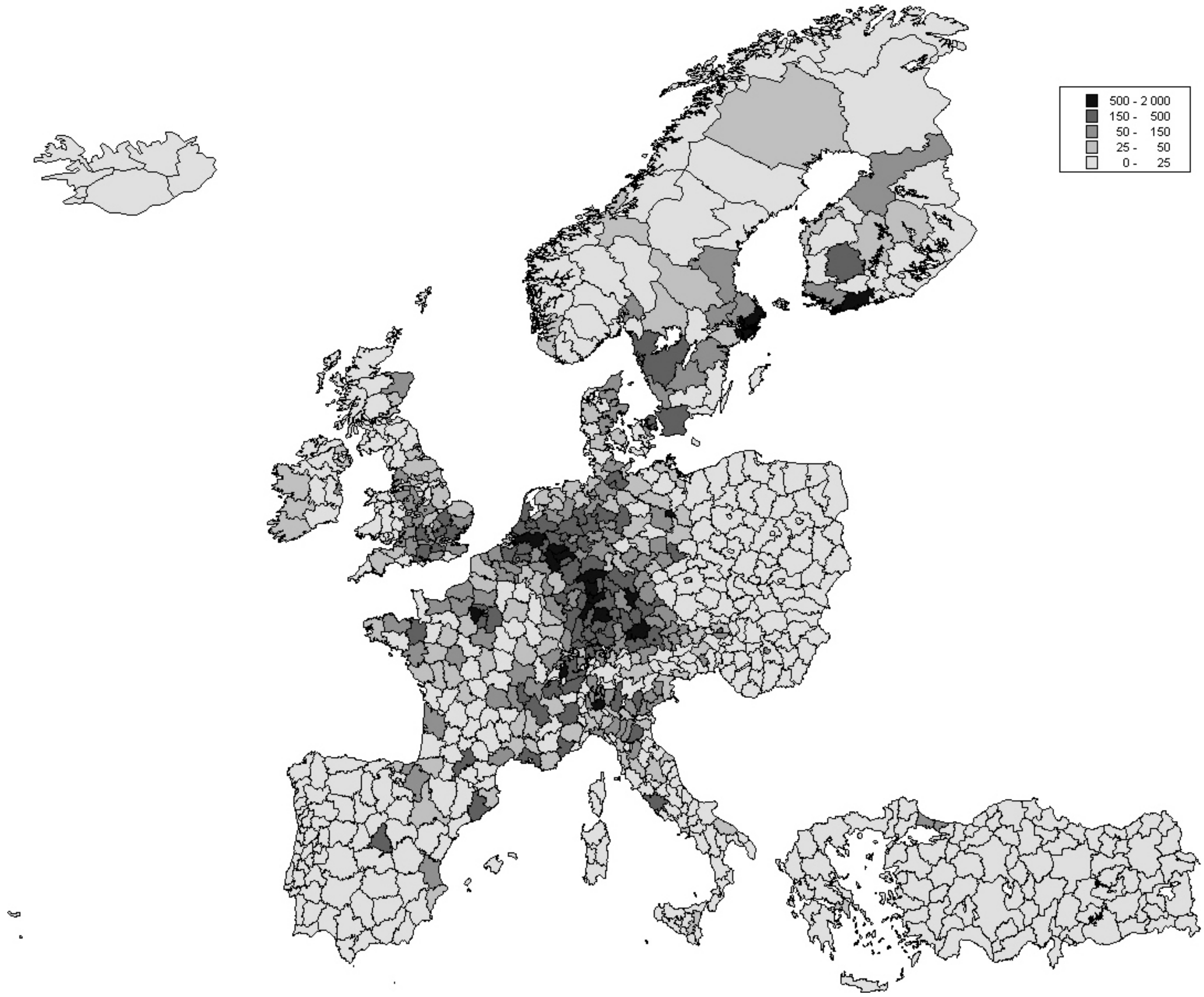


Sources: REGPAT, OECD Patent database and OECD Regional database, February 2008.

Tokyo (485), Osaka (294) and Kyoto (277) are the leading regions of Japan in terms of PCT filings per million habitants. In Korea, the province of Daejeon is the most inventive. In Australia, most inventive activity takes place in the big cities (Sydney, Melbourne, Perth, Adelaide, and Canberra).

Figure 5. Number of patent applications to the EPO, 2004

Europe

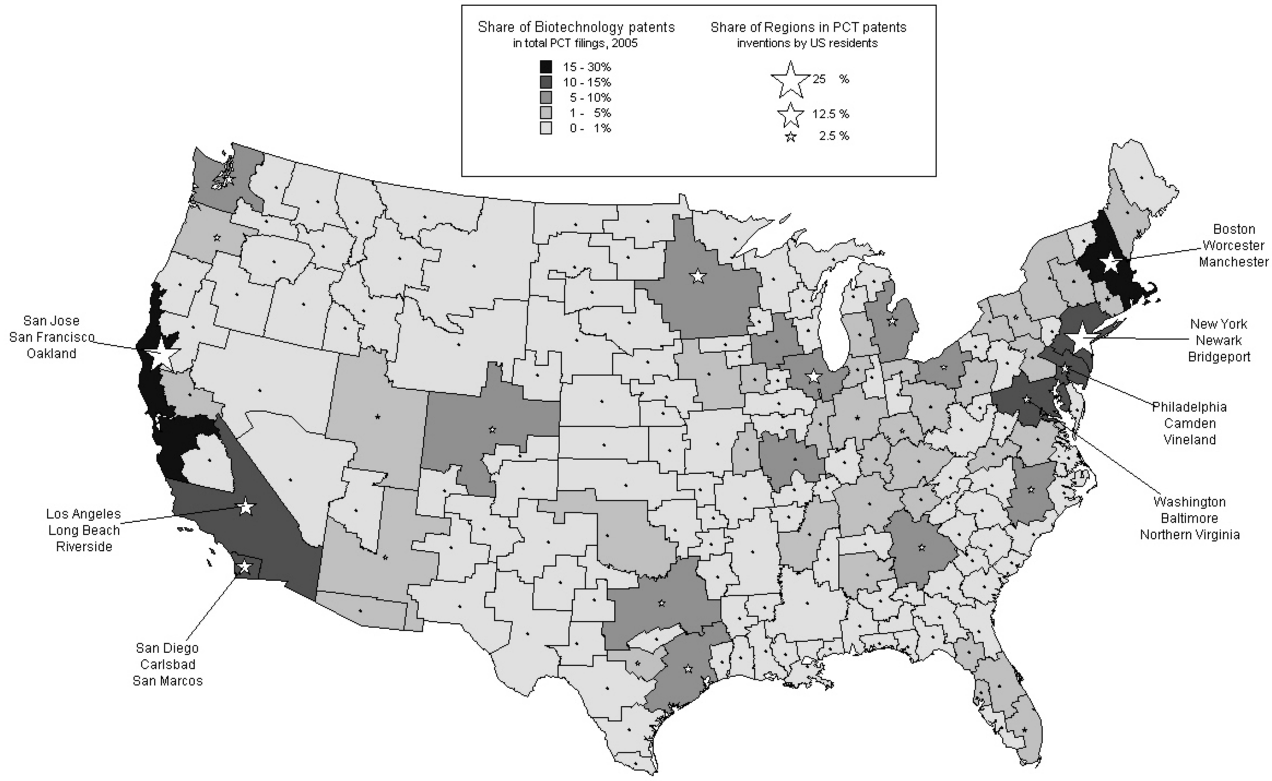


Sources: REGPAT and OECD Patent databases, February 2008.

In 2004, the top seven inventor regions for patent applications to the EPO are non-European regions representing 18% of the total filings: Tokyo (5%), Osaka (2%) and Kanagawa (2%) in Japan, the regions of San Jose/San Francisco/Oakland (3%) and Boston/Worcester/Manchester (1.8%) in the United States and the province of Gyeonggi-do (1.7%) in Korea. Three European regions contribute to more than 1% of total filings each in 2004: Stuttgart (1.6%) and Munich (1.5%) in Germany; Noord-Brabant in the Netherlands (1.5%). Again, the most highly contributing regions of EPO patents in Europe are located on a North/South line crossing Germany, from the Netherlands/Denmark to Austria/Switzerland.

Figure 6. Number of Biotechnology patent applications filed under the PCT, 2005

United States

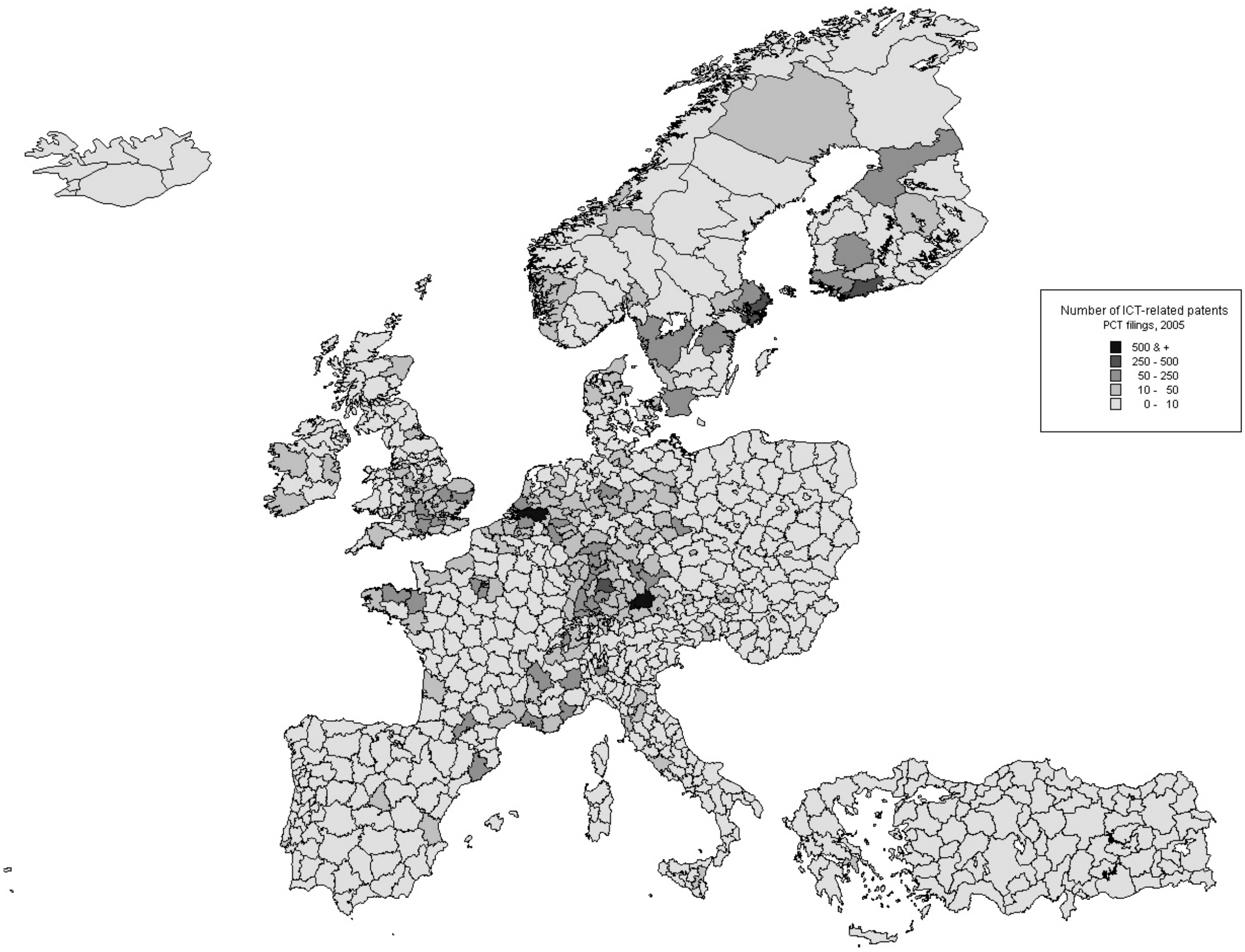


Sources: REGPAT and OECD Patent database, February 2008.

Regions that applied for the largest number of PCT filings in Biotechnology in 2005 are located in the United States and in Japan: In the United States, the largest number of Biotech patent applications originated from either California or the East Coast regions of Boston, New York, Washington and Philadelphia, with between 100 to more than 330 PCT filings in 2005. In Japan, Tokyo (225), Kanagawa (150), Osaka (114), Ibaraki (78) and Kyoto (71) are the principal regions where inventors in biotechnology are located. In Europe, the leading regions are: Berlin (54) and Munich (33) in Germany, Zuid-Holland (36) in the Netherlands, Madrid (33) in Spain, Københavns amt (32) in Denmark and Paris (32) in France.

Figure 7. Number of ICT-related patent applications filed under the PCT, 2005

Europe

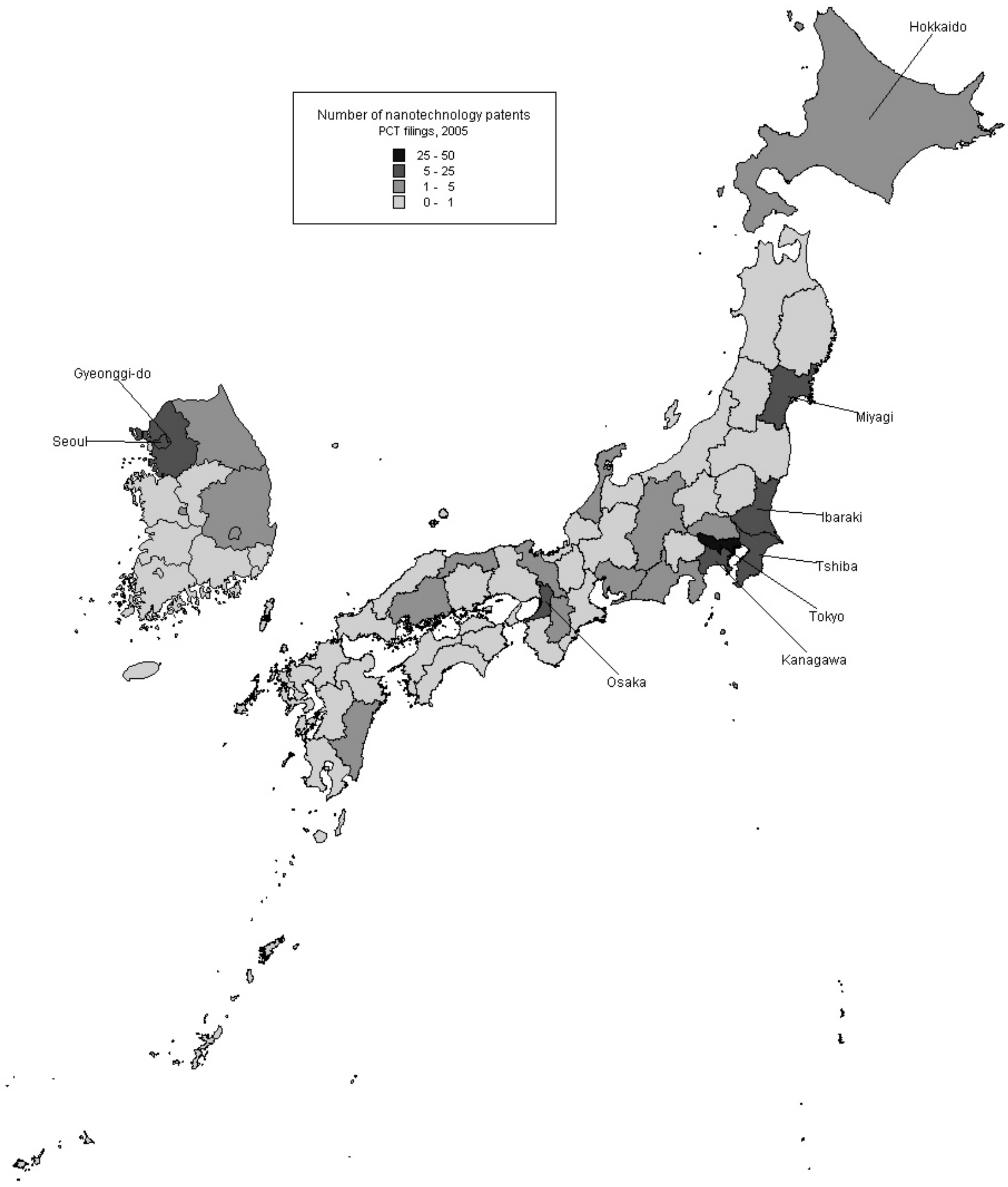


Sources: REGPAT and OECD Patent database, February 2008.

Again, Japan (Tokyo and Kanagawa regions) and the United States (regions of California, Boston, New York and Seattle) are leading in terms of number of PCT filed for ICT-related patents with more than 1 000 patents in 2005. In Europe, the region of Noord-Brabant in the Netherlands follows with 1 038 ICT-related patents. Munich (580), Stuttgart (369) and Aachen (212) are the largest ICT patent-intensive regions of Germany. The leading regions in the Nordic countries are Uusimaa (401) and Pirkanmaa (203) in Finland, Stockholms län (342) and Skåne län (212) in Sweden. In France, ICT-related research activities are conducted by residents living in Paris (291), in Isère (207) and in Hauts-de-Seine (205).

Figure 8. Number of Nanotechnology-related patent applications filed under the PCT, 2005

Japan - Korea

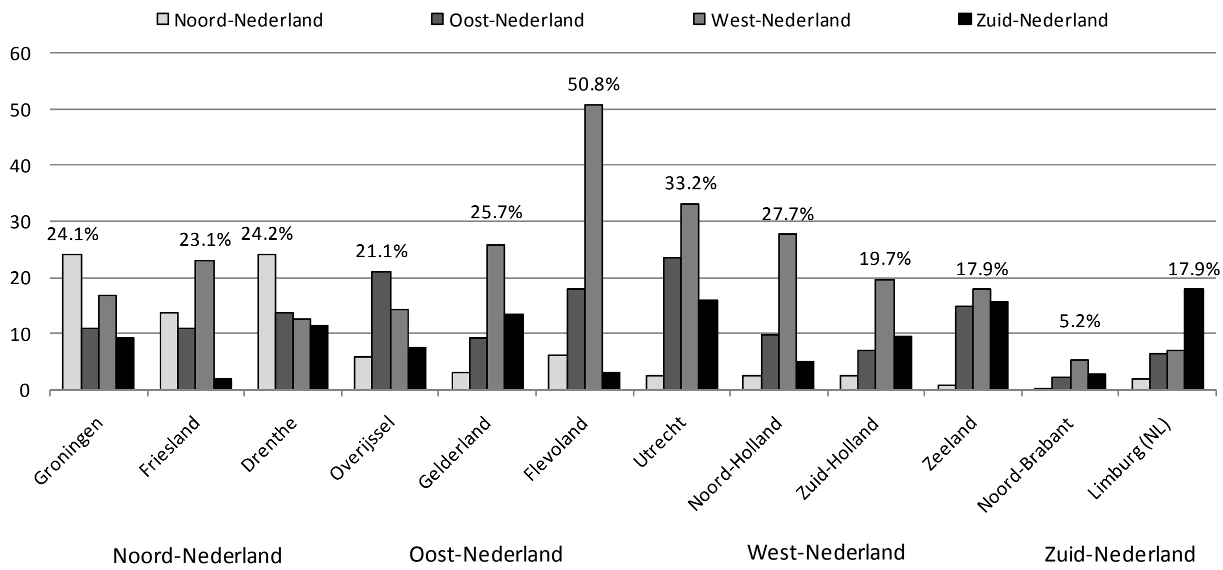


Sources: REGPAT and OECD Patent database, February 2008.

Tokyo (37), Kanagawa (18) and Ibaraki (13) are the first three Japanese prefectures where inventors of nanotechnology-related patents are located. The prefecture of Tokyo ranks in third position worldwide, after the regions of San Jose/San Francisco and Boston in the United States. In Korea, Seoul and Gyeonggi-do province filed respectively 11 and 8 nanotechnology patents under the PCT procedure.

Figure 9. Share of patents with co-inventors from another region, *the Netherlands, 2002-2004*

Intra-Netherlands co-operation



Note: Share of patent applications to the EPO co-invented by at least an inventor from another region.

Sources: REGPAT and OECD Patent database, February 2008.

Regional co-operation figures for the Netherlands (Figure 9 and Table 8) show the influence of the geographical proximity in co-operation between researchers. Most inter-regional co-operation involves inventors from the West-Nederland area. Beyond intra-Netherlands co-operation, some regions tend to collaborate more with foreign regions that are geographically close: Zeeland region co-operate the most with the Flemish region of Belgium. Inventors from the Limburg region co-operate mainly with their closest neighbours: Noord-Brabant, and the German region of Nordrhein-Westfalen. After Germany and Belgium, the largest rates of international co-inventions occur with residents of the United States (Table 8).

Table 8. Share of patents with co-inventors from another region, *the Netherlands, 2002-2004*

Main co-operation

		Belgium	Germany				United States					
		Vlaams Gewest	Baden Württemberg	Bayern	Hessen	Nordrhein Westfalen	California	Connecticut	NewJersey	NewYork	Pennsylvania	Texas
Noord Nederland	Groningen	1.5	1.5	0.0	1.5	0.5	2.0	0.5	0.5	0.5	0.5	0.0
	Friesland	1.4	**3.4**	**3.4**	0.0	0.7	0.0	0.0	0.0	0.0	0.7	0.0
	Drenthe	1.6	0.0	1.6	0.0	1.1	0.5	0.0	0.0	0.0	0.0	0.0
Oost Nederland	Overijssel	1.5	1.2	1.2	0.0	**2.5**	1.5	0.7	1.5	0.7	**2.2**	0.2
	Gelderland	1.6	1.9	0.8	1.2	**5.0**	**2.7**	0.5	1.0	0.7	1.1	0.4
	Flevoland	0.0	**4.7**	0.8	**3.9**	**4.7**	1.6	0.0	0.0	0.0	0.0	0.8
West Nederland	Utrecht	**2.1**	**2.5**	1.7	2.0	**2.6**	**2.8**	1.5	1.5	1.5	2.0	0.2
	Noord-Holland	1.6	1.0	1.7	1.1	1.1	2.0	**2.7**	**2.4**	**2.4**	**2.9**	**3.4**
	Zuid-Holland	1.9	1.6	1.3	0.6	0.9	**2.0**	1.0	1.2	1.2	**2.1**	1.7
	Zeeland	**26.1**	0.7	0.0	0.0	**3.0**	0.0	**2.2**	**3.0**	**3.0**	**3.0**	**7.5**
Zuid Nederland	Noord-Brabant	1.7	0.4	0.2	0.2	1.0	0.5	0.7	0.8	1.1	1.2	0.1
	Limburg (NL)	**7.6**	1.8	1.3	0.7	**13.6**	1.2	0.4	0.5	0.7	1.8	0.1

Note: Share of patent applications to the EPO co-invented by at least an inventor from another region.

Sources: REGPAT and OECD Patent database, February 2008.

4. Conclusion

The availability of regionalised patent data to all students interested in the field should contribute to the development of studies of the regional dimension of innovation. By making regionalised patent data available to all students interested in the field, the OECD aims to stimulate research and contribute to a better understanding of the regional dimension of innovation. In addition, the methodology used for the construction of REGPAT is published, to give users the opportunity to suggest modifications and thus contribute to improvements in the quality of REGPAT. The full technical description of the REGPAT database as accessible to users is provided in annex.

User are encouraged to inform the OECD if they discover possible misallocations of regional codes to certain addresses (correcting false-positives) or if they can provide a regional code where none is provided (correcting false-negatives). Any other feedback that could help improve the quality of REGPAT would be greatly appreciated. REGPAT datasets will be updated right after each release of PATSTAT.

The full technical description of the REGPAT database as accessible to users is provided in annex (Figure A2).

The REGPAT database can be downloaded – please contact Hélène Dernis at OECD: helene.dernis@oecd.org

ANNEX – REGPAT DATABASE DESCRIPTION

The main REGPAT tables used in the regionalisation process are designed to facilitate the links with other tables and datasets:

- Table **ADDRESS** regroups all addresses to be processed with the regionalisation algorithm: *person_id* and *person_address* fields extracted from PATSTAT (EP/US/WO patent documents only); inventors and applicants addresses along with the EP application number from the OECD dataset on EPO filings (EPO direct + Euro-PCT filings at international phase). Consequently, the regionalised addresses can be easily connected to the original data source, as shown in the diagram below (*person_id* for PATSTAT; the pairs *app_nbr/inv_city* and *app_nbr/app_city* for OECD, EPO dataset).
- Table **NUTS3_ADDRESS** summarises the regionalisation procedure, with a specific allocated code (*type_resultat*) for each address analysed (from 0 to 6 – see Figure A1). In case of NUTS allocation with breakdown, the number of NUTS codes is provided (*repartition*). Finally, when the postal code and/or the town name were used to regionalise the address, this information is added in separate fields (*CP* and *ville*).
- Table **NUTS3_ADDRESS_REPARTITION** provides the list of NUTS3 codes allocated to addresses (only *address_id* with a code from 1 to 6 (*type_resultat*) in NUTS3_ADDRESS table), along with the share in case of breakdown (*repartition* from 0 to 100%). The shares can consequently be used in the computation of patent counts by regions, to integrate the relative shares of each NUTS code for fractional counting. Addresses with no region allocated (*type_resultat*=0) are not included in this table.
- Table **REGION** describes the list of NUTS codes and region names (both at level 2 and level 3) by country. A complementary concordance table between NUTS levels and Territorial Levels (TL) can be provided to users upon request.

Organisation of the data for users

The above-described tables were combined to facilitate the use of REGPAT, according to the data sources (Figure A2). Two pairs of tables were created, that can be complemented with the REGION table presented above: Patent applications to the EPO, based on PATSTAT, with regionalised addresses of inventors (EP_INVT_REG) and applicants (EP_APPLT_REG).

- Patent applications filed under the PCT, at international phase (EPO designations), based on OECD dataset, with regionalised addresses of inventors (PCT_INVT_REG and applicants (PCT_APPLT_REG).

Figure A1. REGPAT database diagram

REGPAT main tables

ADDRESS	
ADDRESS_ID	Unique identifier for the address (random)
TYPE_ADDRESS	Data source of the Address* : PATSTAT/OEB_INVT/OEB_APPLT
EXTERNAL_ID	Depending on the source, relating field : Person_id (PATSTAT), App_nbr (OEB_<>)
TYPE_EXTERNAL_ID	Type of external_id for joint tables: Person_id/App_nbr
CTRY_CODE	Address country code - 2 characters
DOC_STD_NAME_ID	*not filled*
PERSON_NAME	*not filled*
ADDRESS	Address field as in data sources : Person_address (PATSTAT) / Inv_city (OEB_INVT) / App_city (OEB_APPLT)
FLG_INC	Binary variable for regionalisation algorithm

* PATSTAT : only EP, US or WO documents
OEB_INVT : EPO inventors / OEB_APPLT : EPO applicants

NUTS3_ADDRESS	
ADDRESS_ID	Address identifier
TYPE_RESULTAT	Regionalisation method**
NBR_REPARTITION	Number of NUTS codes in case of breakdown
CP	Postcode when identified
VILLE	Town name when identified
FLG_REP	Binary variable for regionalisation algorithm

**0-No region allocated
1-Postcode
2-Town name
3-Postcode breakdown
4-Town name after postcode breakdown
5-Unique region (small countries)
6-Mixed postcode breakdown and several town names

NUTS3_ADDRESS_REPARTITION	
ADDRESS_ID	Address identifier***
NUTS3_CODE	Allocated NUTS code
REPARTITION	Share in cases of breakdown (=100 when no breakdown)

*** Only for addresses with valid NUTS code (Type_resultat <> 0)

REGION	
CTRY_CODE	Region country code
NUTS3_CODE	Nuts 3 code
NUTS3_LABEL	Name of region at Nuts 3 level
NUTS2_CODE	Nuts 2 code
NUTS2_LABEL	Name of region at Nuts 2 level

PATSTAT

TLS206_PERSON
person_id
person_ctry_code
doc_std_name_id
person_name
person_address

TLS207_PERS_APPLN
person_id
appln_id
applt_seq_nr
invt_seq_nr

OECD, EPO datasets

EPO_App_names	
App_nbr	EPO application number
App_name	Applicant name
App_code	Applicant code
App_add	Address
App_city	City
App_cou	Country
App_total	Total number of applicants

EPO_Inv_names	
App_nbr	EPO application number
Inv_name	Inventor name
Inv_add	Address
Inv_city	City
Inv_cou	Country
Inv_total	Total number of inventors

Figure A2. REGPAT – User tables

PATSTAT: EPO patents

EP_Applt_reg	
Appln_id	PATSTAT Application identifier
Appln_nr	Patent application number
Publn_nr	Patent publication number
Person_id	PATSTAT person identifier
Address	Address
Reg_code	NUTS3 region code
Ctry_code	Country
Reg_share*	Share ≤ 100
Applt_share**	Applicant share
Reg_type***	Regionalisation method

EP_Invt_reg	
Appln_id	PATSTAT Application identifier
Appln_nr	Patent application number
Publn_nr	Patent publication number
Person_id	PATSTAT person identifier
Address	Address
Reg_code	NUTS3 region code
Ctry_code	Country
Reg_share*	Share ≤ 100
Invt_share**	Inventor share
Reg_type***	Regionalisation method

OECD, EPO database : Euro-PCT at international phase

PCT_App_reg	
PCT_nbr	PCT publication number
PCT_App	PCT application number
EPO_App_nbr	EPO application number
App_name	Applicant name
Address	Address
Reg_code	NUTS3 region code
Ctry_code	Country
Reg_share*	Share ≤ 100
App_share**	Applicant share
Reg_type***	Regionalisation method

PCT_Inv_reg	
PCT_nbr	PCT publication number
PCT_App	PCT application number
EPO_App_nbr	EPO application number
Inv_name	Inventor name
Address	Address
Reg_code	NUTS3 region code
Ctry_code	Country
Reg_share*	Share ≤ 100
Inv_share**	Inventor share
Reg_type***	Regionalisation method

* Region share, when address was allocated to more than one region
** For fractional counts, when more than one applicant/inventor per patent
*** Regionalisation method : 0 - not regionalised ; 1 - postal code ; 2 - town name ; 3 - mixed method