Visual Mining of Science Citation Data for Benchmarking Scientific and Technological Competitiveness of World Countries

Mehmet Can Arslan, Gürdal Ertek[1],
Sabancı University
Faculty of Engineering and Natural Sciences
Orhanlı, Tuzla, 34956, Istanbul, Turkey
mehmeta@su.sabanciuniv.edu
ertekg@sabanciuniv.edu

Abstract

In this paper we present a study where we visually analyzed science citation data to investigate the competitiveness of world countries in selected categories of science. The dataset that we worked on in our study includes the number of papers published and the number of citations made in the ESI (Essential Science Indicators) database in 2004. The dataset lists these values for practically every country in the world. In analyzing the data, we employ methods and software tools developed and used in the data mining and information visualization fields of the Computer Science. Some of the questions for which we look for answers in this study are the following: (a) Which countries are most competitive in the selected categories of science? (i.e. Engineering, Computer Science, Economics & Business) (b) What type of correlations exist between different categories of science? For example, do countries with many published papers in the field of Engineering science also have many papers published on Computer Science or Economics & Business? (c) Which countries produce the most influential papers? This analysis is needed since a country may have many papers published but these papers may be cited very rarely. (d) Can we gain useful and actionable insights by combining science citation data with socioeconomic and geographical data?

Keywords: country, continent, field, papers, citations, citations per paper (cpp), papers per thousand people (pptp), population, visualization

1. INTRODUCTION

Science Citation data is a fundamental tool in measuring and benchmarking academic performance and research capability. It is doubtless that every single paper published in a respective journal in any field of science is a result of serious commitment and endeavor. Nevertheless, producing influential papers which are frequently cited by other researchers is significant in measuring the quality of a published work. In other words, papers having many citations are considered to be

[1] Corresponding author

strong and *influential* ones in their respective fields. Also, these are most likely the papers with major contributions.

A popular criterion throughout the world for measuring scientific quality and productivity is the Science Citation Index (SCI), developed by Thomson Scientific (Thomson Scientific) that lists and tracks the contents of more than 3,700 leading journals in over 100 disciplines. In addition, there is also the SCI Expanded, again maintained by Thomson Scientific, which covers more than 5,800 journals and serves as a broader list.

Within the Evaluation / Analytical subcategory of Thomson Products there exists the Essential Science Indicators (ESI) database, which provides ranking data of journals, scientists, countries and institutions (ESI Database). It covers more than 11,000 journals and 10 million articles from 22 specific fields of research. The database also includes a dataset that lists the number of papers published and cited for each country in these 22 fields, and we have used this dataset in our study.

In this paper, we have visually mined the ESI dataset belonging to year 2004 in order to measure scientific and technological competitiveness of countries around the world. In our analyses, we have primarily benefited from the advantages of information visualization for deriving new insights.

In Section 2, we describe the related literature on the use of visual data analysis to investigate science citation data. In Section 3, we give a brief description of the data as well as introducing the readers to the field of information visualization. In Section 4, we explain the features and capabilities of the software tools that we have used in our study. In Section 5, we introduce our four main research questions and search for answers to them. We include concluding remarks following each of our analyses corresponding to each question. Finally in Section 6, we summarize our findings and propose new paths for expanding the study presented here.

2. RELATED LITERATURE

Today, thanks to the internet, one can easily find extensive amount of data on country statistics. The traditional sources of such data have been CIA World Factbook (CIA), United Nations (UN), WDI (WDI), and OECD (OECD). Now, the data in these resources can be found online not only in the websites of these organizations but also in large data repositories, such as NationMaster (NationMaster). On the analysis side, online services such as Many Eyes (Many Eyes) and Swivel (Swivel) enable analysis of a wide variety of data from within a web browser. As of April 2007, Many Eyes stores ~70,000 data sets and offers 14 different types of visualizations, whereas Swivel stores ~3,300 data sets and offers only a few types of visualizations.

The GapMinder software (GapMinder) developed by the Sweden based GapMinder Foundation is a remarkable demonstration of how colored scatter plots and other visualizations can be used to derive interesting - and sometimes highly unexpected - insights. GapMinder is designed to especially support the analysis and benchmarking of world countries and allows users to import new

dimensions and recent statistics into existing data files. One striking feature of GapMinder is that it can animate the change of dimension values of countries throughout time and export these animates in the popular Adobe Flash format.

Some of the academic work related to our study is summarized below:

The work of Ingwersen et al. (2001) is the paper that we have found to be closest to our study, since the authors employ visualization, particularly bar charts and colored scatter plots for measuring and benchmarking research performance of different countries. Their data mainly consists of number of papers and number of citations for seventeen OECD countries in two four-year time intervals in nine social science fields. They propose a new performance measurement metric (indicator), namely Tuned Citation Impact Index (TCII). Three hypotheses are proposed and tested through visual and statistical analysis.

Ashton and Sen (1989) US and International companies based on patent data regarding sodium-sulfur batteries. The data includes dimensions such as number of patents, number of inventors, areas of emphasis and patent citations.

Boyack et al. (2002) present how the VxInsight software tool that they developed can be used in creating landscape visualizations that reflect the relationships between scientific articles. The visualizations show domains of science as hills and articles as glyphs. This tool is especially helpful in determining science and technology management strategies by managers.

Erickson (1996) questions the validity conclusions drawn in patent citation studies. The author focuses on the telecommunications equipment industry and uses data of six different firms in his analyses. He concludes that the differences in patent systems and other factors make it impossible to directly compare the patent data of USA, Japan and Europe.

Leta (2005) presents a snapshot of the scientific productivity of Brazil by analyzing the patterns of publications in astronomy, immunology and oceanography in comparison to the overall publication pattern of Brazil. In our study, we focus on comparison of different countries whereas Leta (2005) focuses on only a single country. The methodology of the author can be applied as a second level of analysis for the countries of interest identified through our methodology.

3. DATA AND METHOD

3.1 Dimensions in the Dataset

First, we list the dimensions in the dataset that we use, mostly taken from Essential Science Indicators (ESI) database. Figure 1 gives a snapshot of the original data obtained from ESI. The dimensions marked with a star (*) are derived from other resources (Wikipedia, WDI Online) and included in the analysis. The dimensions marked with (**) are computed using data in other dimensions.

Country	:	name of the country
Continent **(*)**	:	continent that the majority of the Country's land In geographical means belongs to
Field	:	broad name of the branch of science
Papers	:	number of papers published in 2004 in the Field specified, according to the ESI database
Citations	:	number of citations in 2004 that refer to the papers published in 2004 or earlier coming from that Country
Citations per Paper (CPP) **(**)**	:	ratio of Citations to Papers for that Country
Population **(*)**	:	population of that country (Most of the data is gathered from World Development Indicators (WDI) database and belongs to 2005 while others are retrieved from Wikipedia and commonly belong to 2006 or 2007)
Papers per Thousand People (PPTP) **(**)**	:	for a specific country, the ratio of the Papers to Population multiplied by 1000; will be referred to as PPTP

Country	Field	Papers	Citations	Citations per Paper
ALGERIA	PHYSICS	...	...	...
ALGERIA	CLINICAL MEDICINE	183	2,377	12.99
ALGERIA	MOLECULAR BIOLOGY & GENETICS	...	...	...
ALGERIA	MATERIALS SCIENCE	...	...	...
ALGERIA	ENGINEERING	...	...	...
ALGERIA	GEOSCIENCES	...	...	...
ALGERIA	ENVIRONMENT/ECOLOGY	...	...	...
ALGERIA	SPACE SCIENCE	...	...	...
ALGERIA	NEUROSCIENCE & BEHAVIOR	...	...	...
ALGERIA	MATHEMATICS	...	...	...
ALGERIA	COMPUTER SCIENCE	...	...	...
ALGERIA	BIOLOGY & BIOCHEMISTRY	...	...	...
ALGERIA	CHEMISTRY	...	...	...
ARGENTINA	CLINICAL MEDICINE	5,820	44,956	7.72
ARGENTINA	PHYSICS	...	...	...
ARGENTINA	PLANT & ANIMAL SCIENCE	...	...	...
ARGENTINA	NEUROSCIENCE & BEHAVIOR	...	...	...
ARGENTINA	MOLECULAR BIOLOGY & GENETICS	...	...	...
ARGENTINA	SPACE SCIENCE	...	...	...
ARGENTINA	ENVIRONMENT/ECOLOGY	...	...	...
ARGENTINA	MICROBIOLOGY	...	...	...
ARGENTINA	GEOSCIENCES	...	...	...
ARGENTINA	ENGINEERING	...	...	...
ARGENTINA	IMMUNOLOGY	...	...	...
ARGENTINA	PHARMACOLOGY & TOXICOLOGY	...	...	...
ARGENTINA	MATERIALS SCIENCE	...	...	...
ARGENTINA	MATHEMATICS	...	...	...

Figure 1. A snapshot from the original ESI Dataset

One of the major challenges regarding the interpretation of the data throughout our study was to decide which dimension is key in answering a particular

question. For instance, even though the number of papers published (Papers) is an important metric for a country, the number of citations (Citations) is also an important indicator of the influence of papers published. At this point, Citations per Paper may seem like a reasonable dimension for evaluation. However, since a country may have few papers and citations and still have a bigger CPP value than another country with considerable number of papers and citations, CPP was not a complete solution, either. For instance, as seen in Figure 1, in Clinical Medicine field, **Argentina** dominates **Algeria** both in Papers and Citations and still has a lower CPP value than **Algeria**. There are several examples of such occurrences. Hence, we have tried to make a comprehensive analysis by using all the metrics simultaneously in different visualization settings. Also for a socioeconomic analysis, we have added two separate fields; namely, the Population of the countries and the Papers per Thousand People (PPTP).

For each research question, we tried to give answers including all of the selected fields, i.e. Engineering, Computer Science, and Economics & Business. However, in some cases, we have just focused on the outputs with the most interesting insights and sometimes only on the Engineering Field as a representative. In addition, we have filtered the data meaningfully where necessary to obtain better or focused insights. Lastly, if a data record was found to include an extremely deviated value for a specific dimension (for example, the CPP value), then that data records was omitted from the visualization so that the remaining points could be distinguished from each other.

3.2 Information Visualization

Information visualization is the expanding field of Computer Science that enables data mining through visually representing data. The goal is to provide the analyst with visualizations through which hidden patterns in the data can be revealed, as aimed in statistical techniques and other data mining methods. Favorably, most of the visualization applications come with understandable forms of user interfaces. Hence, information visualization is much more powerful in extracting new insights when a high level of human perception and judgment is involved.

Butz and Schmidt (2005) refer to information visualization as allowing people to draw the right conclusion out of a presentation. Well-known sources of information visualization include books by Tufte (2001) and Shneiderman (1999). As the field of information visualization is expanding, the number of available tools (i.e. the dedicated software packages as described in the next section) is also increasing. Plaisant (2004) gives a structured method of how to benchmark information visualization tools.

This project focuses on the comparison of countries according to the Science Citation Data belonging to selected fields and analyses are based on pre-defined research questions. For each question, separate visualization methods are suggested and utilized in accordance with the nature of the research question. Even though there may be several alternative visualization methods for displaying the same data, we have used our experience and judgment in selecting only the most insightful and the ones that are easy to interpret. Moreover, relevant graphical tools from statistics (ex. histograms) were also utilized when necessary.

4. SOFTWARE TOOLS

4.1. Mondrian

Mondrian is a freely available data analysis software package with rich features. It is developed at RoSuDa (RoSuDa) using the Java programming language (Java). It mainly reads tab-delimited data files as input. However, it can also retrieve data through a direct database connection (Mondrian). The tool supports MS Windows, Linux and Macintosh platforms (RoSuDa) and we have used the MS Windows XP version in our study. Mondrian currently has Mosaic Plots, Scatter Plots, Maps, Bar Charts, Histograms, Parallel Coordinates/Box Plots and Box Plots y by x implemented under an integrated whole (Mondrian). In this paper, we have used Mondrian for obtaining Parallel Coordinate Plot visualizations.

4.2 Miner3D Enterprise

Miner3D Enterprise (Miner3D) is a powerful software for visual data mining. The software is used in a variety of application areas ranging from chemistry to finance, biology to marketing, etc. It provides an interactive environment for the users for advanced data analysis and decision support. Miner3D Enterprise currently has Scatter2D, Scatter3D, Tile, LineGraph, Heat, Bar2D, Bar3D, Clustering, Factor Analysis and Self Organizing Maps (SOM) implemented. It also provides basic statistical analysis including mean, modus, standard deviation, and others. Full compatibility with MS Excel and MS Access is a major advantage of the software, making it an appropriate choice for supporting data analysis tasks carried out in spreadsheets. In this paper, we have used Miner3D for obtaining Scatter2D (colored scatter plots) visualizations.

4.3 Omniscope

Omniscope is a data analysis software package by the firm Visokio (Visokio). It is a viable alternative for visual data analysis. Because of its user-friendly and colorful interfaces, different visualization capabilities, spreadsheet (MS Excel) and database (MS Access) connectivity and its diversified filtering options, it is a suitable software for usage. It provides 10 pre-built data views: Chart, Table, Graph, Tile, Pie, Bar, Map, Pivot, Tree and Portal views. It is also possible to generate customized visualizations in Omniscope by combining readily available data views. For the same dataset, the user is able to display several views on the screen at the same time, dynamically change filtering conditions and sort data with respect to any attribute. The free version of Omniscope is limited to import up to 500 rows (records) and 6 columns (dimensions) from MS Excel. In our analyses of the ESI data with Omniscope, this limitation did not cause any problems since the number of dimensions we needed were always less than six at a time. In this paper we have used Omniscope for obtaining Tile and Map visualizations.

4.4 Stata

Stata (Stata) is a complete statistical analysis software with advanced data management and graphics capabilities. It can run under MS Windows, Linux and Macintosh platforms. One can run commands from the command line or use the menus in the graphical user interface (GUI). The statistical language of Stata is demonstrated extensively under the Help menu and is easy to learn. Below, there is an example of the necessary command lines to import data from MS Excel, which we have used many times in this study. The ↵ sign denotes the "Enter" key on the keyboard. Stata has applications in almost all areas of science. In our analyses, we have benefited from Stata 9 in plotting the histograms of values in several dimensions of the data.

```
clear ↵
odbc load, dsn("Excel Files;DBQ=c:\stata_files\eng.xls") table("Sheet1$") ↵
```

5. ANALYSES

Preliminary Findings

Table 1. General Overview of the ESI dataset

Total Papers	Total Citations	Total Distinct Fields	Number of Listed Countries	Total Number of Records
9,762,529	90,337,998	22	145	1856

- The 41 countries having published papers within all 22 fields are Argentina, Australia, Austria, Belgium, Brazil, Canada, Chile, Czech Republic, Denmark, England, Finland, France, Germany, Greece, Hungary, India, Ireland, Israel, Italy, Japan, Mexico, Netherlands, New Zealand, North Ireland, Norway, Peoples Republic of China, Poland, Portugal, Russia, Saudi Arabia, Scotland, Slovakia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Turkey, USA and Wales (in alphabetical order).

- The Field with the biggest number of countries involved is Clinical Medicine with 103 countries having published in this field in 2004.

- The number of countries having published in the Engineering Field in 2004 is 93.

- The number of countries having published in the Computer Science. Field in 2004 is 75.

- The number of countries published in Economics & Business Field in 2004 is 75.

Having described the fields of the ESI dataset in the previous sections and stated the preliminary findings above, now we answer the primary research questions that we posed in the Abstract, along with the related analyses.

5.1. Competitiveness in Different Fields of Science

In this section, we search for the answer to the following research question: Which countries are most competitive in the selected fields of science, namely Engineering, Computer Science and Economics & Business?

We have carried out the first analysis on citation data for the Engineering Field (Shortly ENG). In the 2D Scatter plot (colored scatter plot) analysis (Figure 2) carried out with Miner3D, the x-axis shows Citations, y-axis shows Papers, and the color shows CPP. Both the Papers and the Citations values for **USA** are extremely high compared to other countries in all fields and thus we have excluded **USA** from the visualizations in Figures 2, 3, 4, 5, 6 and 7.

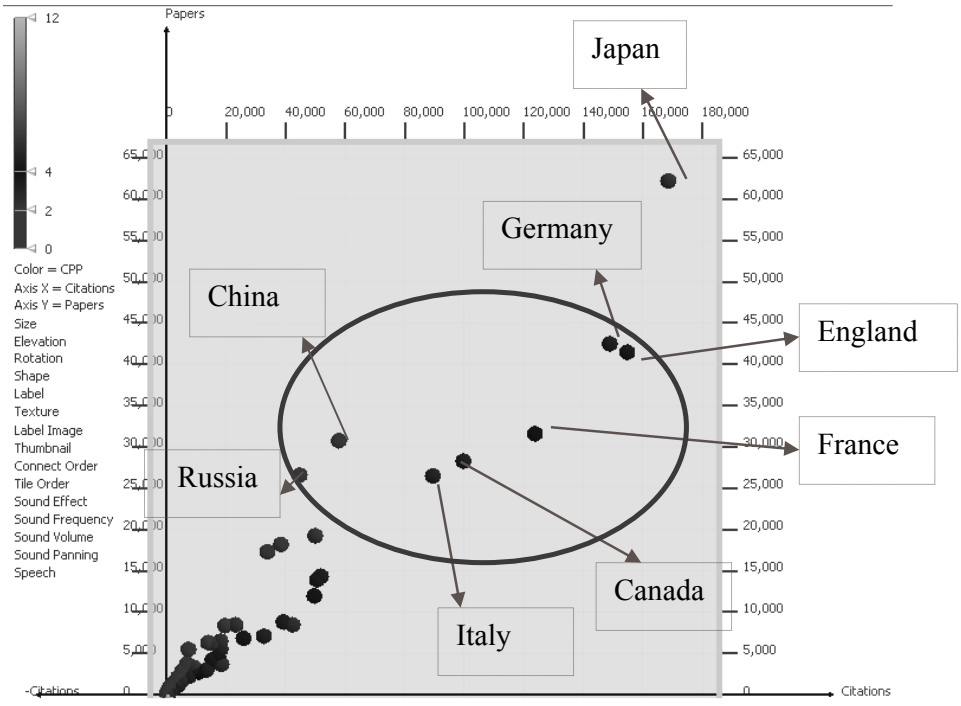

Figure 2. Analysis of the Engineering Field (**USA** is excluded)

Figure 2 clearly shows that there is almost a linear pattern between Citations and Papers among the listed countries. **Japan** is the dominating figure in this category following **USA**. **England** and **Germany** are very close to each other in terms of both Citations and Papers. **France, China, Canada, Italy** and **Russia** are the other leading countries in this field in terms of Papers. However, we can exclude **China** and **Russia** from the previous list if we consider Citations (or CPP as well) instead of papers. These 8 countries together with **USA** can be considered as the most competitive players in the Engineering Field.

One thing to emphasize on Figure 2 is the axes ranges. The greatest value on the x-axis is 180,000 and the greatest value on the y-axis is 65,000 (excluding **USA**'s values). However, in the other figures for Computer Science (shortly CS) and Economics & Business (shortly E&B) fields, we will see much smaller numbers on both axes. This is a result of how fields were constructed in the ESI database. Engineering involves many sub-disciplines in itself, however, the other fields covered in this paper (i.e. CS and E&B) stand on their own.

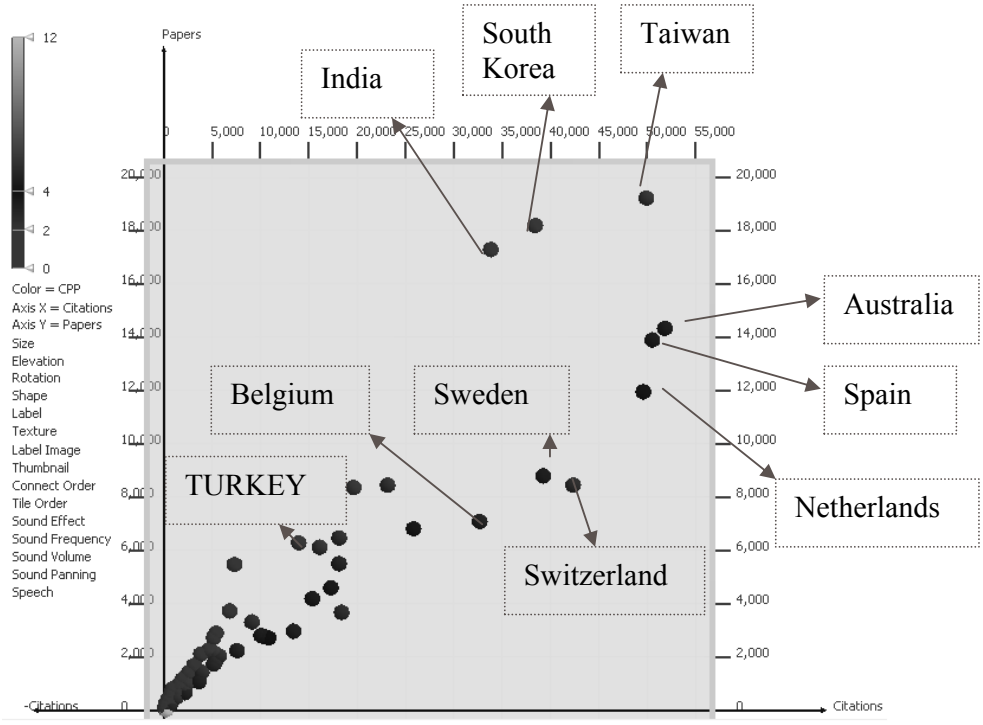

Figure 3. Analysis of the Engineering Field (**USA** and the Countries labeled in Figure 2 are excluded)

If we focus on the remaining unlabeled countries in Figure 2 by zooming in, we obtain Figure 3. Now, we have new players at the top corner: **Taiwan, South Korea** and **India** with the most number of papers and **Australia, Spain** and **Netherlands** with the most number of citations. Following them, we see **Switzerland, Sweden** and **Belgium.**

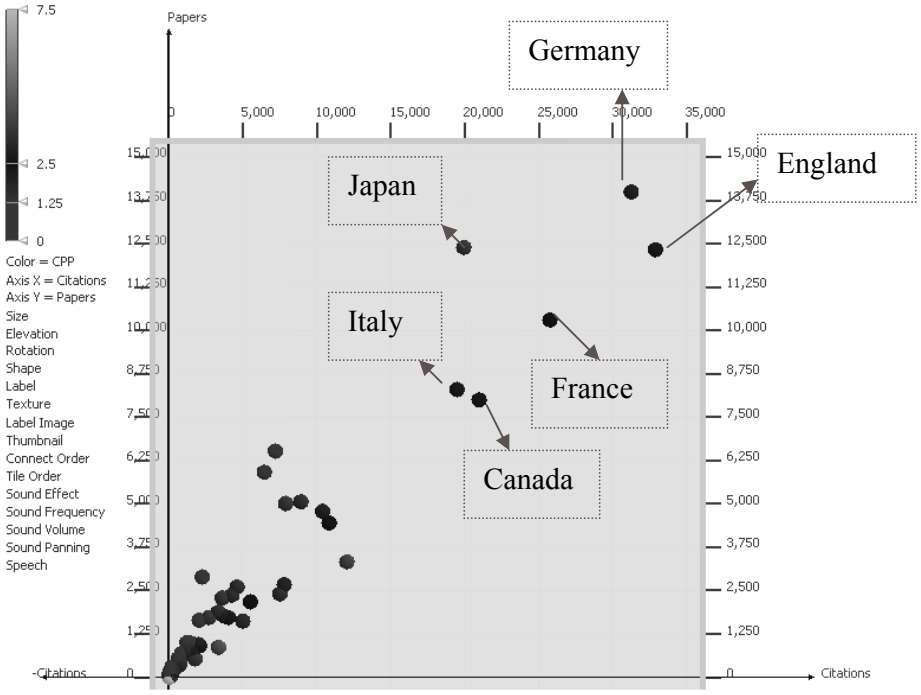

Figure 4. Analysis of the Computer Science Field (**USA** is excluded)

Figure 4 presents a colored scatter plot of the Computer Science Field. Here, we see a slightly different pattern than what we observe in Figure 1. Here, the leading players of the Field have a more remarkable separation and their order is different. In terms of Citations, **England** comes first, following **USA**. This indicates that the scientific papers produced in **England** have higher impact compared to other countries (except **USA**). However, regarding Papers, we see that **Germany** is more competent than **England**. **Italy** and **Canada** are very close to each other. **France** leads **Italy** and **Canada** with approximately 5,000 more citations and also more papers. **Japan** is very close to **England** in terms of Papers but has just about the same number of Citations as **Italy** and **Canada** have. Apart from the key players, the distribution of the rest of the Field is quite similar to the distribution of the rest of the countries in the plots for Engineering.

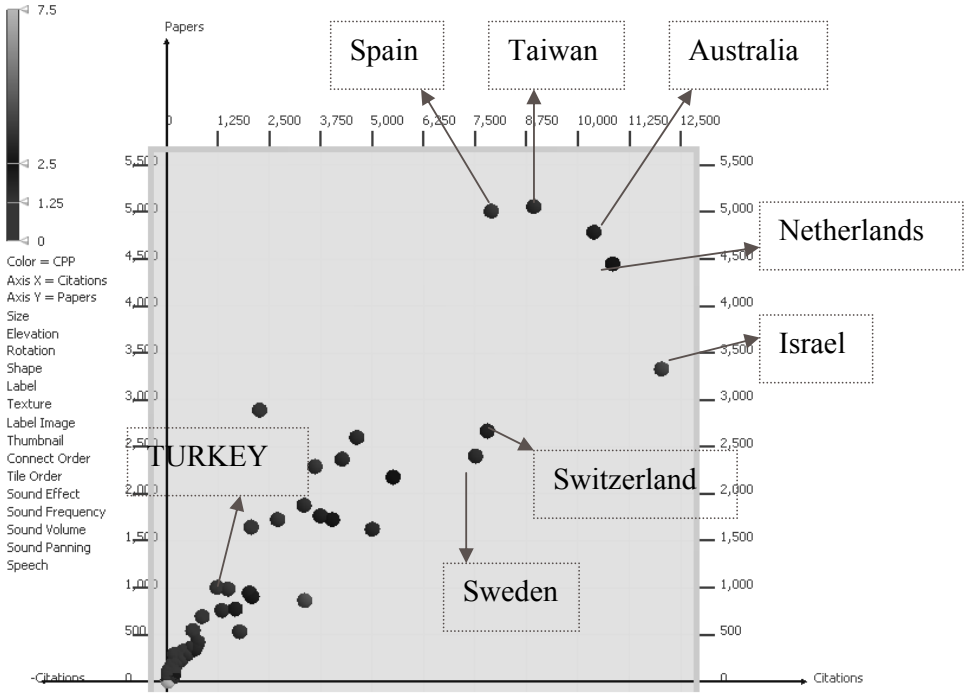

Figure 5. Analysis of the Computer Science Field (**USA** and the Countries labeled in Figure 4 are excluded)

When we zoom in Figure 4 to observe the countries not labeled in Figure 4 as in Figure 5, we see **Taiwan** and **Spain** with the most number of papers and **Israel** with the most number of citations. **Australia** and **Netherlands** are close to each other with respect to both metrics. **Switzerland** and **Sweden** stand close to **Taiwan** and **Spain** in terms of Citations whereas they are far behind regarding Papers.

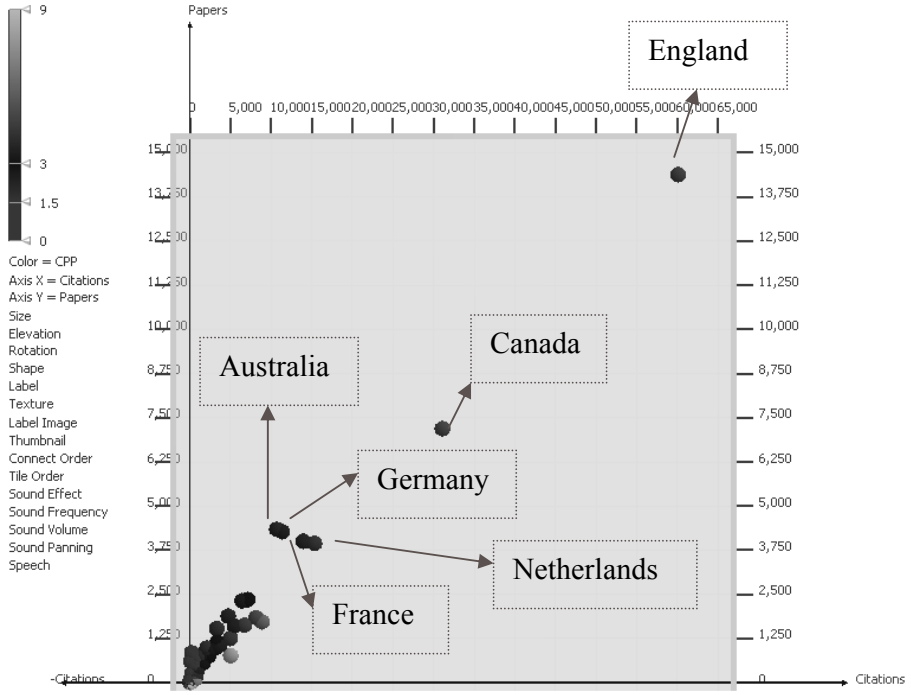

Figure 6. Analysis of the "Economics & Business" Field (**USA** is excluded)

Figure 6 presents a colored scatter plot of the Economics & Business field. Here, we observe that there are significant changes in the competitiveness of countries when compared to the previous two fields. As we still observe a linear pattern, the higher values of the axes are dominated by a single Country, **England** (**USA** is excluded). We observe that there is a much greater gap between **England** and other countries in the Economics and Business Field compared to the two fields mentioned earlier. This suggests that **England** is much more superior in this Field to other countries. The third most competitive Country (following **USA** and **England**) in this Field seems to be **Canada**, however, her Papers and Citations values are nearly half of those of **England**. Following these countries, we observe that **Australia, Germany, France** and **Netherlands** are positioned in very close proximity to each other.

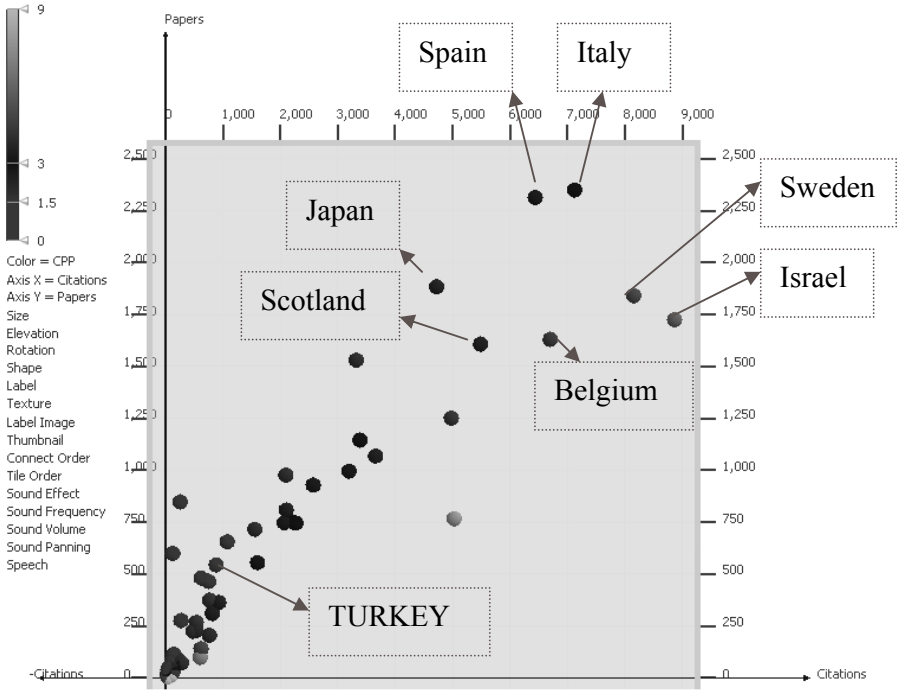

Figure 7. Analysis of the Economics & Business Field (**USA** and the Countries labeled in Figure 6 are excluded)

When **USA** and the countries labeled in Figure 6 are removed from the data, we obtain Figure 7. In this secondary group of countries in the Economics & Business field, **Spain** and **Italy** stand at the top with respect to Papers, whereas **Sweden** and **Israel** lead in terms of Citations. **Japan** is a bit better than **Sweden** and **Israel** with respect to her Papers value, but behind them with respect to her Citations. **Belgium** and **Scotland** are two other competitive countries in this Field.

Concluding Remarks

i. Key Players of Each Field

Figures 2-7 plotted in Miner3D give us an understanding of the competitiveness of the world countries in the selected fields of science. As seen in these figures, the leading countries in each field differ from each other, whereas there are some countries recurrently appearing as leaders in all three fields (**USA, Germany, England, France,** and **Canada**).

One can devise a scheme to classify the countries in terms of Academic Productivity and Popularity (**APP**) according to the Papers and Citations values. We suggest the definition of the following four classes:

1) **Dynamic and Competent** (Countries with many papers and many citations)
2) **Dynamic but Incompetent** (Countries with many papers but few citations)
3) **Static but Competent** (Countries with few papers but many citations)
4) **Static and Incompetent** (Countries with few papers and few citations)

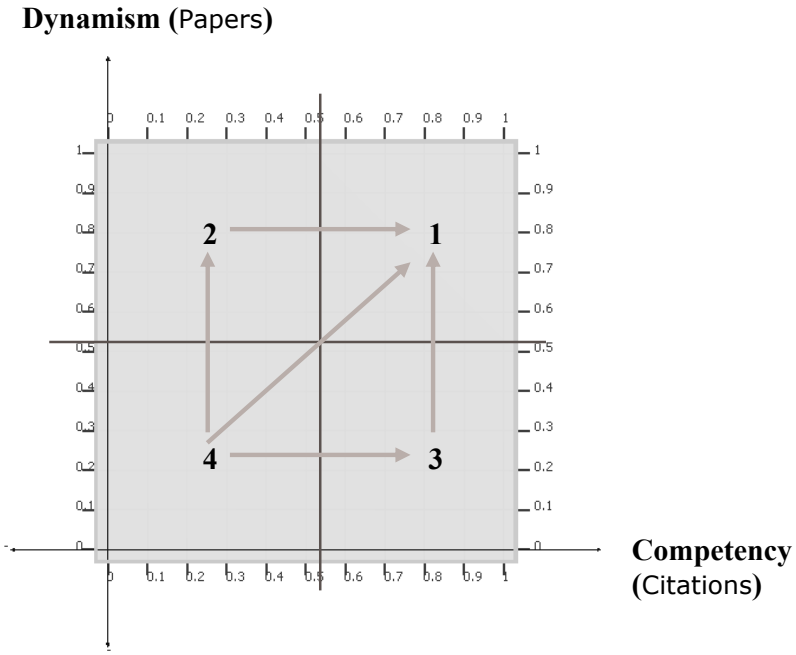

Figure 8. APP Classification

Figure 8 shows a visualization of the suggested APP classification scheme. The green arrows denote the possible target paths for the countries belonging to each class. For example, a 4th category country must either try to publish more papers to jump into category 2 or publish more qualified papers to move into category 3 or do them both simultaneously to move into category 1. A country in already in category 3 for a specific Field should primarily target producing more papers to be in the 1st category, on the other hand.

Actually, the APP classification that we propose here is self-explanatory. By using Figures 2-7, each Country can be classified into the corresponding class. Then, the target of all the countries for that specific Field of science is clear.

ii. Turkey's Position in These Three Fields:

We see **Turkey**'s position in Figures 3, 5 and 7 where we zoom in (when we exclude the best players of the Field). In all of these figures, **Turkey** appears in the leftmost tail of the pattern even after the best countries are excluded from the visualizations.

Although **Turkey**'s position is a little bit better in Engineering Field compared to in Computer Science and Economics & Business, it is clearly seen that she stands in the 4th category in all three fields of science. Hence, she has to progress further to be a key player in terms of academic dynamism and competency.

5.2. Correlations Between Different Fields of Science

In this section, we search for an answer to the following questions: What type of correlations exists between different fields of science? For example, do countries with many published papers in the field of Engineering also have many papers published on Computer Science or Economics & Business?

For this analysis, we tried to compare the values for the Papers field of the common 41 countries that have papers published in all fields. For this visualization, we have used the parallel coordinate plot option of Mondrian. In Figure 9, the blue lines denote the selected countries and grey lines denote the unselected ones. Here, we have selected the leading group of Engineering in terms of Papers as described in Figure 2 previously. (**1: Russia, 2: China, 3: Japan, 4: Italy, 5: Germany, 6: France, 7: England, 8: Canada**; **USA** again excluded due to the high deviation)

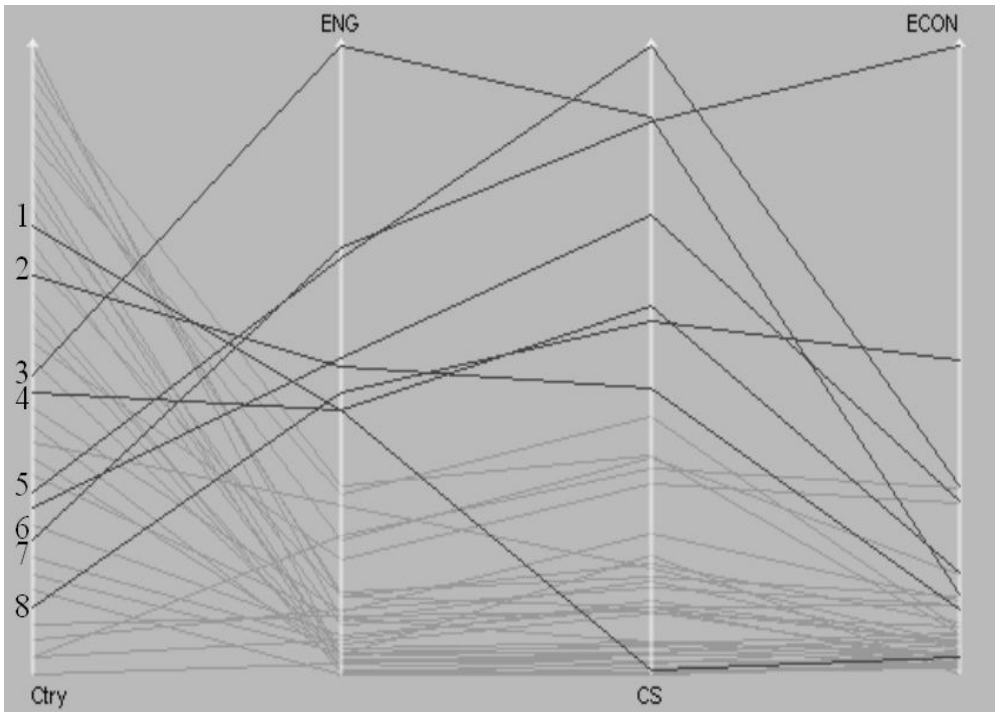

Figure 9. Correlations between Different Categories of Science (Papers) (**USA** excluded)

As seen from the general distribution of the lines in Figure 9, the selected fields can be considered as almost positively correlated, meaning that a Country that is productive in Engineering Field is likely to be productive in Computer Science and Economics & Business fields as well. However there are certain exceptional countries. For example, **Russia** (1) is among the leading players in Engineering Field but not in Computer Science or Economics & Business fields. **China** (2) is good in Engineering and Computer Science but not in Economics & Business. **Japan** (3) also has a pattern similar to **China**.

Germany (5), **France** (6), **England** (7), and **Canada** (8) are better in CS Field than they are in ENG. **England** and **Canada** are also very competent in E&B whereas **Germany** and **France** follow behind.

Remarks

i. Suggested Collaborations for Turkey:

According to the results from the previous analyses on competency and correlations, below are our suggestions on academic collaborations for **Turkey**:

Suggestions for Increasing Productivity

Engineering: **Japan, Russia** and **China** from Asia; **England, Germany, France, Italy** from Europe; **Canada** and **USA** from America (also see Figure 2)

Computer Science: **Japan** from Far East; **England, Germany, France** and **Italy** from Europe; **Canada** and **USA** from America (also see Figure 4)

Economics and Business: **England, Germany, France** and **Netherlands** from Europe; **Australia** from Australia; **Canada** and **USA** from America (also see Figure 6)

Suggestions for Increasing Quality

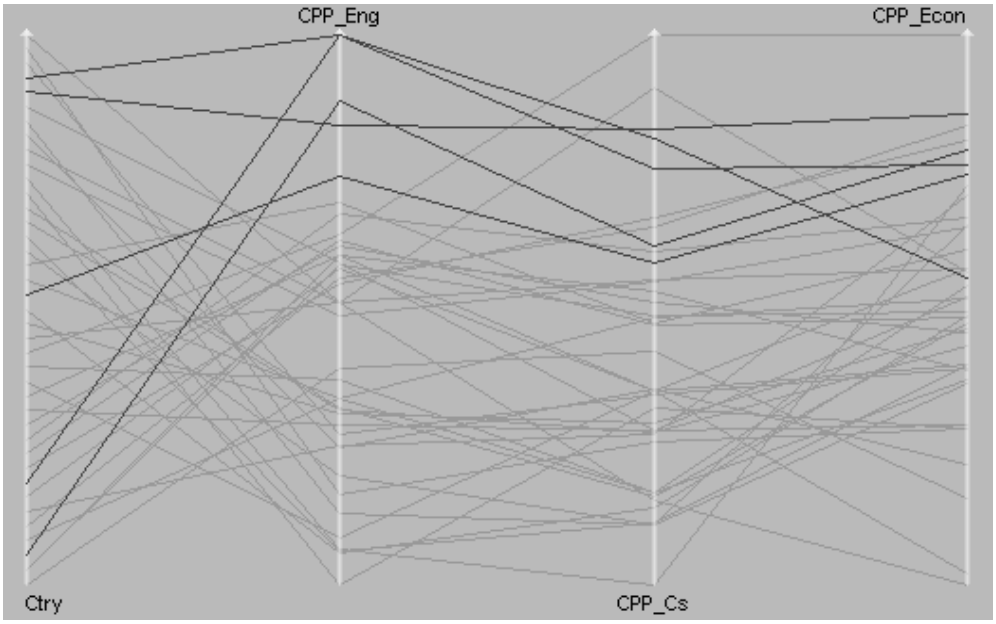

Figure 10. Correlations between Different Categories of Science (Citations) (**USA** excluded)

In terms of producing more qualified papers; on the other hand, **Turkey** should collaborate with countries producing the top quality (the most cited) papers. For the Engineering Field, **Switzerland, Sweden, Netherlands, Denmark** and **Belgium** have the greatest CPP values as seen in Figure 10. **Switzerland, Sweden** and **Denmark** also have better values of CPP in Computer Science Field than Belgium and Netherlands. **Sweden, Belgium and Switzerland** also appear in the leading edge in Economics & Business Field regarding CPP. Hence, collaboration with these European countries presents great opportunities for **Turkey** in improving her academic publication quality. Also in Section 5.4, a geographical interpretation of suggested collaborations is included.

5.3. Countries with the Most Influential Papers

In this section, we search for an answer to the following question: Which countries produce the most influential papers?

To answer this question, we have used another visualization scheme, namely the Tile View in Omniscope.

A *partial* answer to this question can be obtained by analyzing the CPP values of the countries for each Field. The answer is only partial, since CPP alone is not a direct measure of academic influence, as discussed in Section 3.1. (Figure 1) However, CPP can still tell a lot when combined with another metric, as we have applied in this section. In Figures 11, 12 and 13, the color shows CPP and the tile size shows Papers. Green color denotes higher values of CPP where red color denotes lower values and yellow denotes the mid-values.

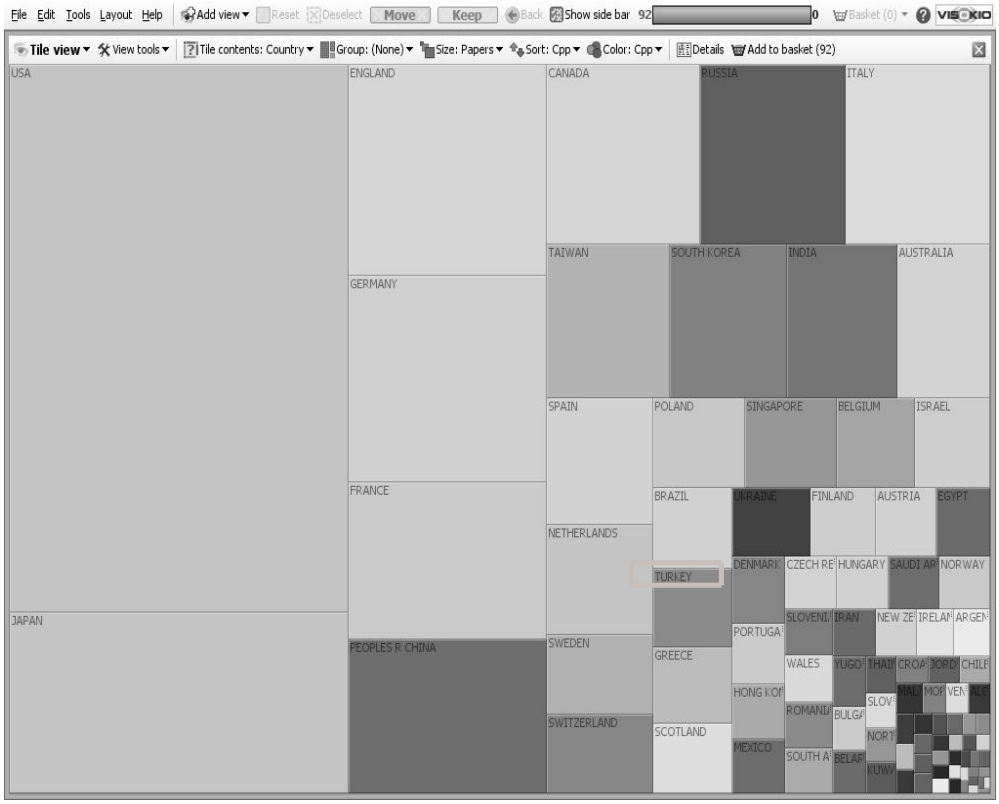

Figure 11. Analysis of Engineering (ENG) Field with a focus on CPP (**Monaco** excluded)

Figure 11 illustrates that among the key players of Engineering Field (based on their Papers values), most of the countries are also academically influential, as indicated by the dominant green color. (**USA, England, Germany, France, Canada**) However, **Japan, China** and **Russia** are not much influential as suggested by the yellowish and red colors in their tiles. Among the remaining countries, **Switzerland, Denmark, Belgium** and **Hong Kong** seem to be the ones with the most influential papers and **Ukraine** seems to be the worst.

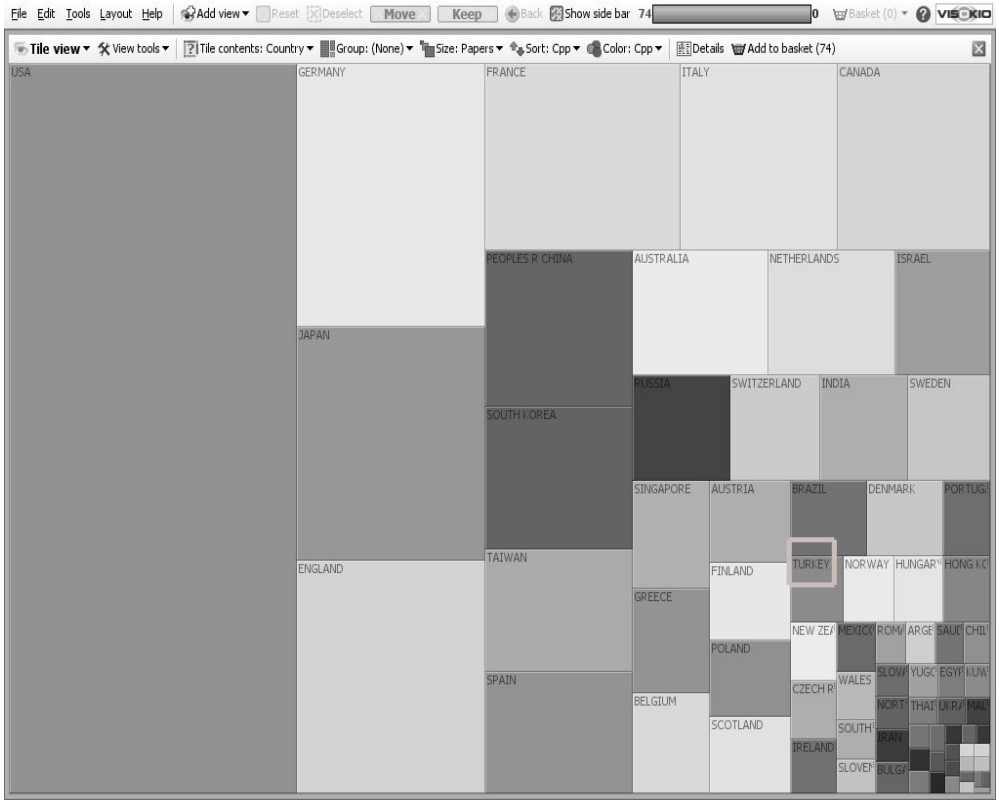

Figure 12. Analysis of Computer Science (CS) Field with a focus on CPP (**Liechtenstein** excluded)

Figure 12 presents a visualization of the number of papers (Papers) and CPP in the Computer Science (CS) field. Liechtenstein is excluded from the visualization, since it has a very high CPP value. Its presence would distort the colors in the figure and prevent us from gaining deeper insights. If we analyze Figure 12, **USA** grasps attention at first sight with her dominating tile size (high Papers value) and intense green color (high CPP value). In CS field, we observe much more orange, yellow and red colors for countries with higher Papers values (ex. **Germany, Japan, China** and **South Korea**) compared to the earlier tile visualization in Figure 11 for Engineering Field. **USA, England, France** and **Canada** still have a green color meaning that they are also influential in Computer Science Field like they are in Engineering. Among the remaining countries with medium tile size, **Russia** appears to be the least influential (as can be read from the dark red color of its tile) and **Israel** seems to be the most influential (as can be read from the dark green color of her tile).

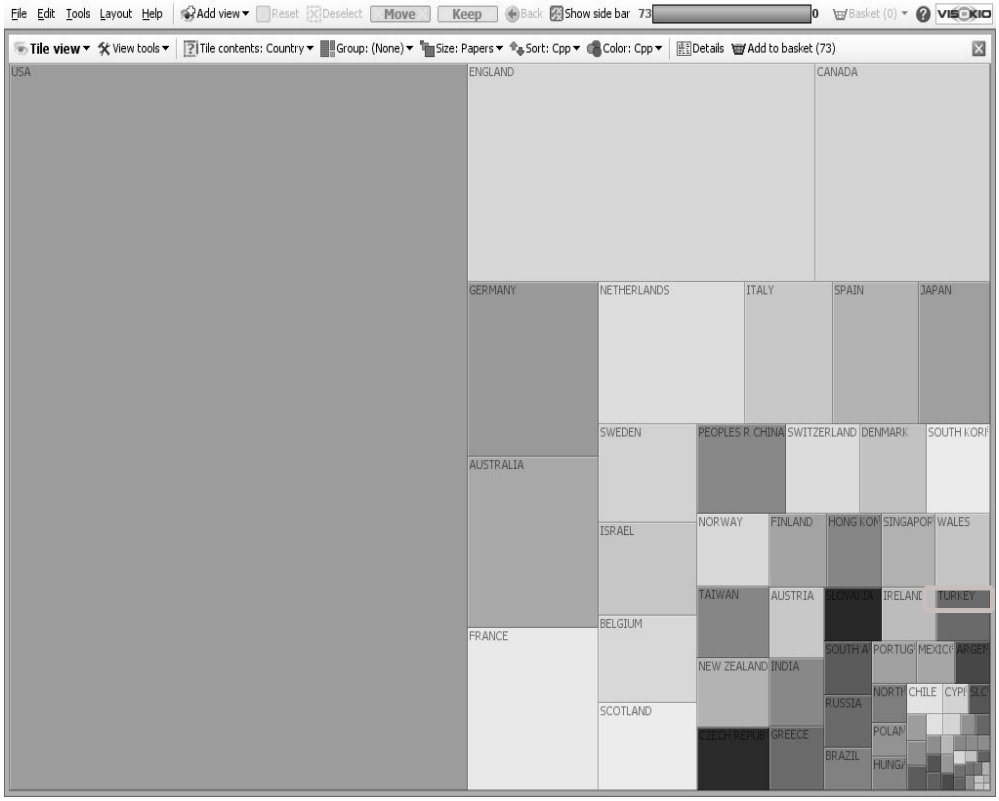

Figure 13. Analysis of "Economics & Business" (E&B) Field with a focus on CPP (**Macao** and **Ecuador** excluded)

Figure 13 presents a visualization of the Papers and CPP in the Economics & Business Field. **Macao** and **Ecuador** are excluded since they have very high CPP values. In Figure 13, we again observe **USA** dominating the Field, as can be interpreted from the large size and dark green color of her tile. We observe that the dominance of **USA** is even more in this Field compared to the earlier two fields, since her tile occupies a much larger area in the visualization this time. **Germany, Australia,** and **France** have many papers published, however with little influence. **England** and **Canada**'s papers have more significant influence. Apart from **USA**, the majority of the tile area is covered with non-green colors and **Czech Republic** and **Slovakia** seem to be the least influential countries (as can be read from the dark red colors of their tiles). **Hong Kong** and **Israel** seem to be the most outstanding countries among the ones having similar tile sizes (as can be read from the dark green colors of their tiles).

Concluding Remarks

i. Influence on a "Field" Basis:

If we consider the total tile area and dominant green color in Figure 11, we can conclude that the majority of the papers produced in Engineering Field have a high influence (high CPP value) compared to the maximum CPP for any country. Figure 14 also supports this claim statistically. When we analyze the histogram plotted using Stata, we see that more than %50 of the countries have a CPP values greater than 2 and more than %30 have CPP values greater than 3. The histogram

in Figure 14 shows that the CPP value for Engineering follows a right-skewed distribution. However, the distribution does not follow a smooth pattern so it would be difficult to fit a well-known statistical distribution to the data.

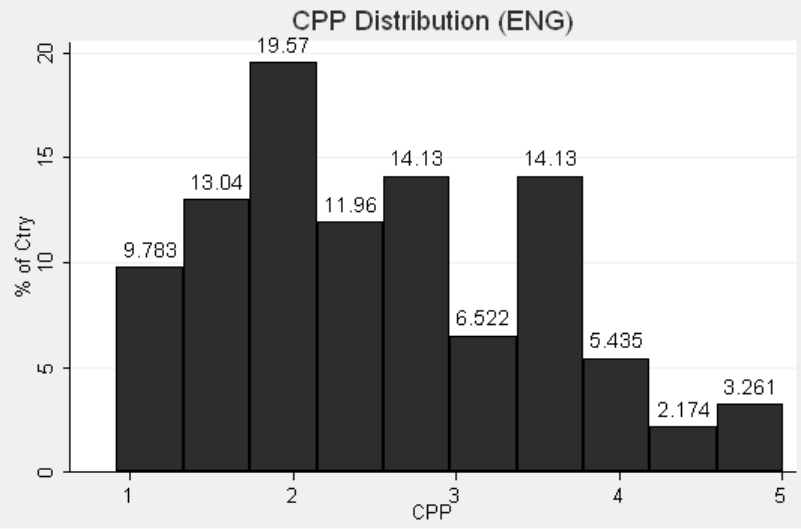

Figure 14. CPP Distribution of Engineering (**Monaco** is excluded)

On the other hand, in Computer Science Field, we observe orange, yellow and red colors more than the green in the tile area, meaning that the CPP values of the majority of the countries fall well behind the maximum CPP value (Figure 12). We can verify this pattern further by observing the histogram plotted for CS field for analyzing the CPP distribution (Figure 15). According to the histogram for CS field, more than % 60 of the countries has a CPP value less than 2 and nearly % 20 of them have a CPP value even less than 1. The distribution of the CPP value for CS is right-skewed, as in the CPP for Engineering, but this time the histogram follows a much more smoother pattern. This suggests that one can fit a well-known statistical distribution to the data with a statistical software package.

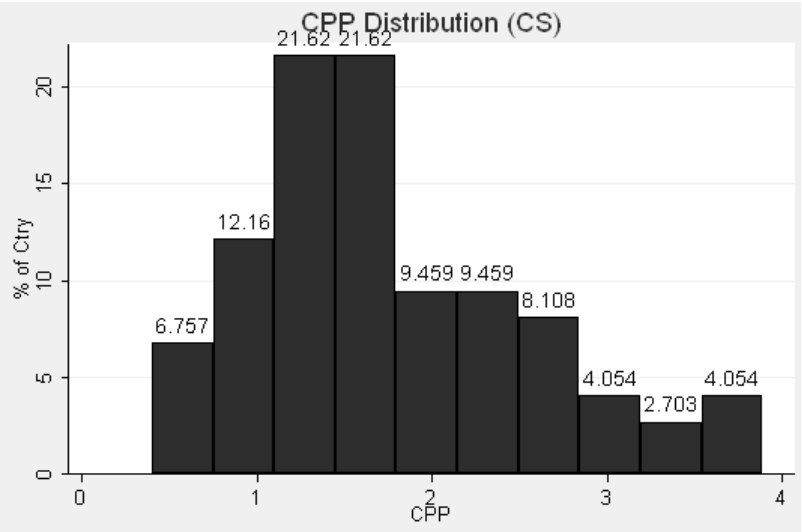

Figure 15. CPP Distribution of Computer Science (**Liechtenstein** is excluded)

Finally, the Economics & Business Field is also a Field in which a great percentage of countries produce influential papers, as indicated by the large green area in Figure 13. Figure 16 suggests that nearly % 75 of countries have CPP values greater than 2 in Economics and Business Field. It can be seen that the distribution of CPP values in this field exhibits an almost-symmetric pattern, suggesting that one can consider fitting a normal distribution to this data.

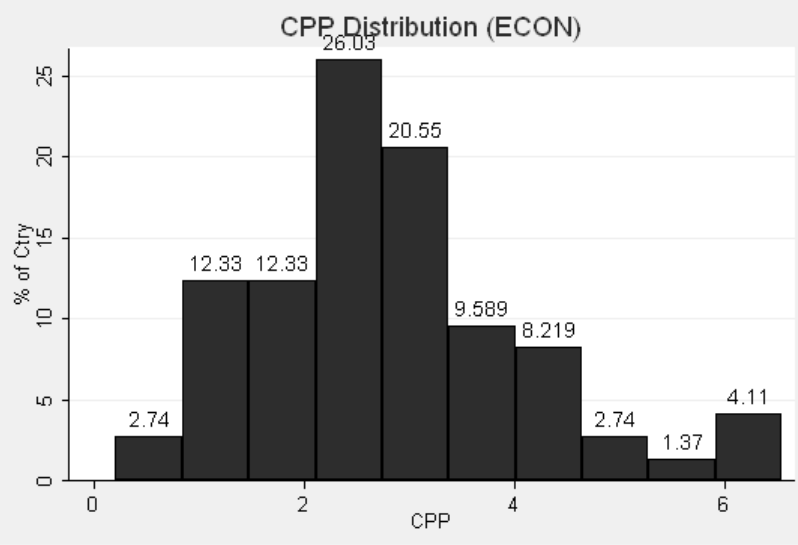

Figure 16. CPP Distribution in Economics & Business (**Macao** and **Ecuador** excluded)

ii. Influence on a "Continent" Basis:

It also makes sense not just to look at individual countries but also the continents they belong to. In Figures 17, 18 and 19, we aim at gaining an understanding of influence on a continental basis by again using the Tile View in Omniscope. In these figures, color represents the CPP values and tile sizes are set equal for each Country.

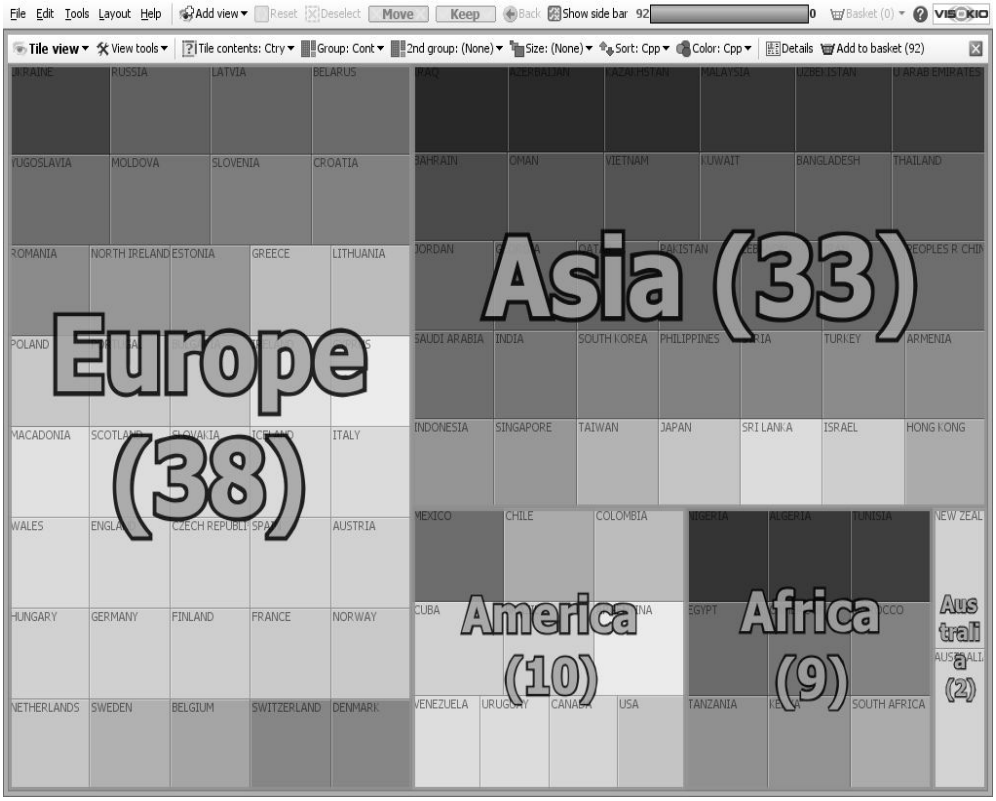

Figure 17. Distribution of CPP in Engineering with respect to Continents (**Monaco** is excluded)

In Engineering Field (Figure 17), Europe leads with respect to CPP. More than half of European countries have a shade of green color meaning that their published papers in this Field are influential compared to the others. The continent of America has less number of countries (10) than Europe (38) but also displays a pattern similar to Europe regarding the citation influence. Asia seems to be the least influential continent with almost all of its tiles in red or orange tones. Australia only has 2 countries listed and they seem to be influential. We see Africa with 9 countries that all have a very small influence in terms of CPP, with **South Africa** performing the best.

In Computer Science (Figure 18), again with the largest number of countries (36), Europe comes at the beginning. However, the influence (as can be read from the overall color of the tiles in Europe) of the European countries in this Field is not as much as in Engineering. The continent of America has 9 countries publishing papers in the Field of Computer Science and their influence is similar to that in Engineering. Asia has far less number of countries than it has in Engineering (22) and seems to follow a similar pattern in terms of influence as before. However, the most influential countries belong to Asia in the Field of Computer Science (**Hong Kong** and **Israel**). From the 2 countries of Australia, Australia seems to become less influential and New Zealand seems to be more influential in this Field compared to Engineering. Africa also has less number of countries compared to what it has in Engineering (5) and they are not influential as it is obvious from their dark red colors.

Figure 18. Distribution of CPP in Computer Science with respect to Continents (**Liechtenstein** is excluded)

In Economics & Business Field, we observe less number of countries in Europe (29) publishing papers compared to the previous fields (Figure 19). Europe is not very influential in E&B, as it is apparent from the orange and red tones of the tiles. Again as implied by the colors of the tiles, Asia and even Africa seems to be more influential than they were in the Engineering and Computer Science fields. Africa also has more number of countries in this field compared to Computer Science. (10) America does not differ much with **USA** still being very influential and Australia as a whole seems to be less influential than it was both in Engineering and Computer Science.

Figure 19. Distribution of CPP in Economics & Business with respect to Continents (E&B) (**Macao** and **Ecuador** are excluded)

i. Turkey's Influence in Academic Citation:

According to Figure 11, which shows the Engineering Field, **Turkey** seems to be less influential than the countries with similar tile sizes, such as **Brazil**, **Greece** and **Scotland**.

In Figure 12, which shows the CS Field, **Turkey**'s tile size (Papers value) is almost the same as the tile sizes of **Norway**, **Hungary** and **Hong Kong,** but all of these three countries are more influential than **Turkey.**

Finally, in Figure 13, which depicts the Economics & Business Field, when compared with the countries with similar tile sizes, **Turkey** seems to be close to **Greece** and **India**, better than **Slovakia**, and worse than **Ireland** and **Austria** with respect to influence (CPP value).

To summarize, **Turkey** should enhance her position in these 3 fields of science in terms of influence of her papers, which is measured through CPP values in our study.

5.4. Combining Science Citation Data with Socioeconomic and Geographical Data

In this section, we search for an answer to the following question: Can we gain useful and actionable insights by combining science citation data with socioeconomic and geographical data?

One can indeed obtain interesting insights by linking research outputs with geographical and socioeconomic characteristics the countries. From this respect, we focused on gathering a basic understanding of such relations regarding 41 selected countries of the world. These are the countries that produce papers in all 22 fields of science according to the ESI dataset. (Mentioned in the preliminary findings section) For practical purposes, we have given a name to these 41 countries; referring to them as the countries on the "**Productivity Belt"**.

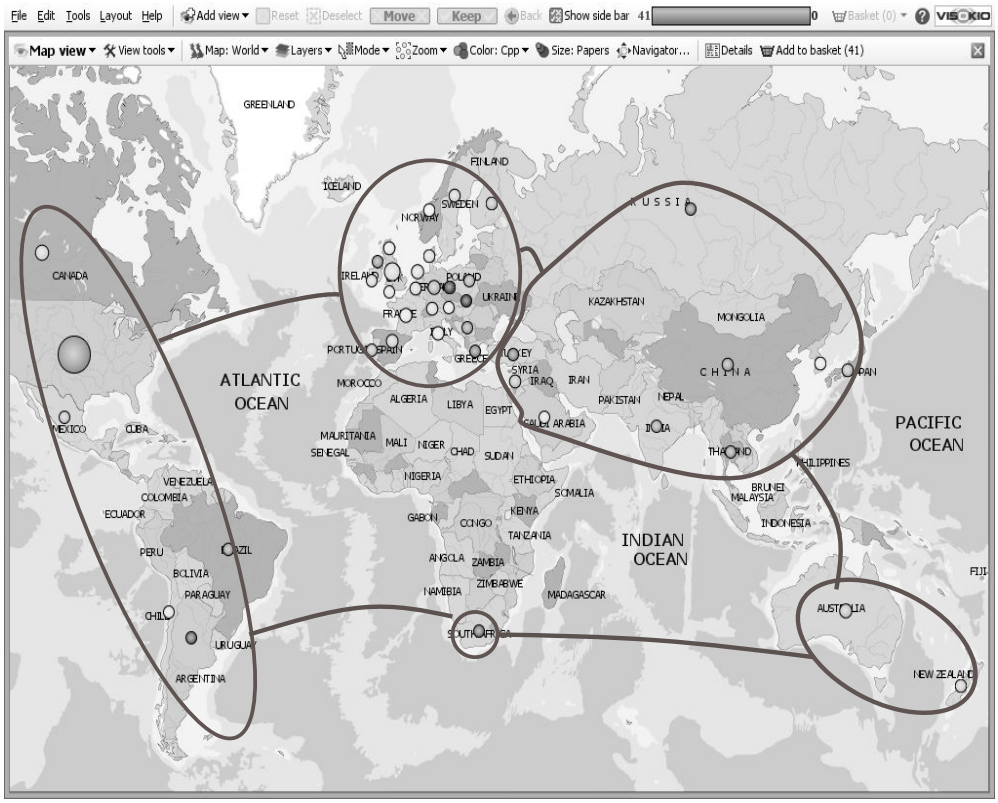

Figure 20. Productivity Belt (Engineering)

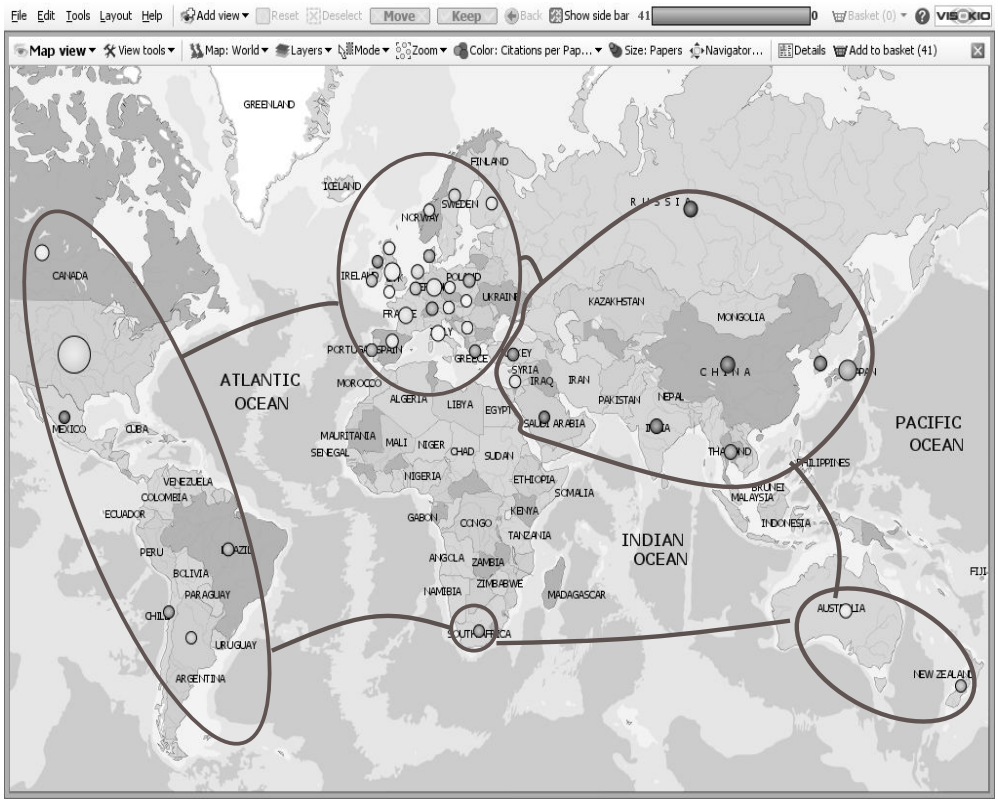

Figure 21. Productivity Belt (Computer Science)

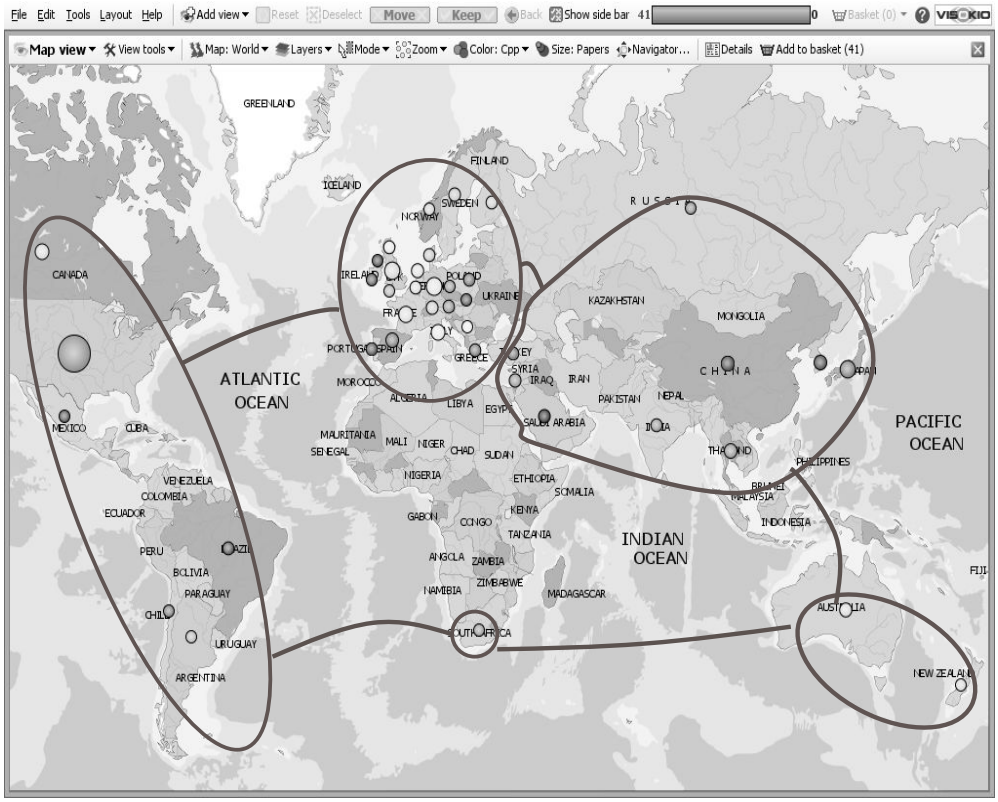

Figure 22. Productivity Belt (Economics & Business)

In these visualizations (Figure 20, 21 and 22), sizes of the colored circles (which we will refer to as "glyphs") show Papers and color shows CPP. Red colors denote lower values of CPP and green colors denote higher values. Yellow color represents mid-values.

As can be observed immediately, Europe definitely dominates the productivity belt with the largest number of countries included. (23 out of 41) Also, most of these countries produce influential papers, as can be observed from their colors in tones of green (Figure 20, 21 and 22).

The continent of America also has a high impact in terms of Papers and is included on the productivity belt with 6 countries. **USA** is consistently the top player in the world in terms of academic productivity in all fields.

Africa's "academic poverty" is another remarkable insight. The only African country that resides on the productivity belt is **South Africa**. We can easily conclude that this scene is an accurate reflection of Africa's political and economic instability that has been continuing for decades.

Asia has 9 countries on the productivity belt. Among the three fields, Asia's color and size position differ a lot and its effect in the productivity belt can not be neglected.

Australia (Oceania) has 2 countries on the productivity belt, namely **Australia** and **New Zealand**. In terms of CPP, both of them sit in the mid-range. In terms of Papers, they are not very powerful, as well.

Turkey is in the middle of the ranges in all three fields when number of papers (Papers) is considered. However, as her color indicates, she does not produce highly cited papers compared to the countries with approximately the same number of papers.

Figures 23 and 24 show a zoomed view of Europe, as positioned on the productivity belt. The only difference between the two figures is that in Figure 23, size of the glyphs shows Citations values whereas in Figure 24, it shows Papers values.

In both Europe figures, **Turkey**'s glyph has intentionally been included to allow the reader combine the previous academic comparisons regarding Turkey with her geographical position (Specifically for comparing Turkey with European countries).

One insight provided by Figure 23 and 24 is the identification of countries which Turkey can collaborate with around her neighborhood. **Greece**, the closest country to Turkey, has red color, indicating that the papers published by **Greece** are not very influential. On the other hand, **Hungary** and **Italy**, still being very close to Turkey have more influential papers published. When the whole Europe is considered, **Switzerland, Denmark, Belgium** and **Sweden** stand out with their influential research, so Turkey should consider collaborating with these countries.

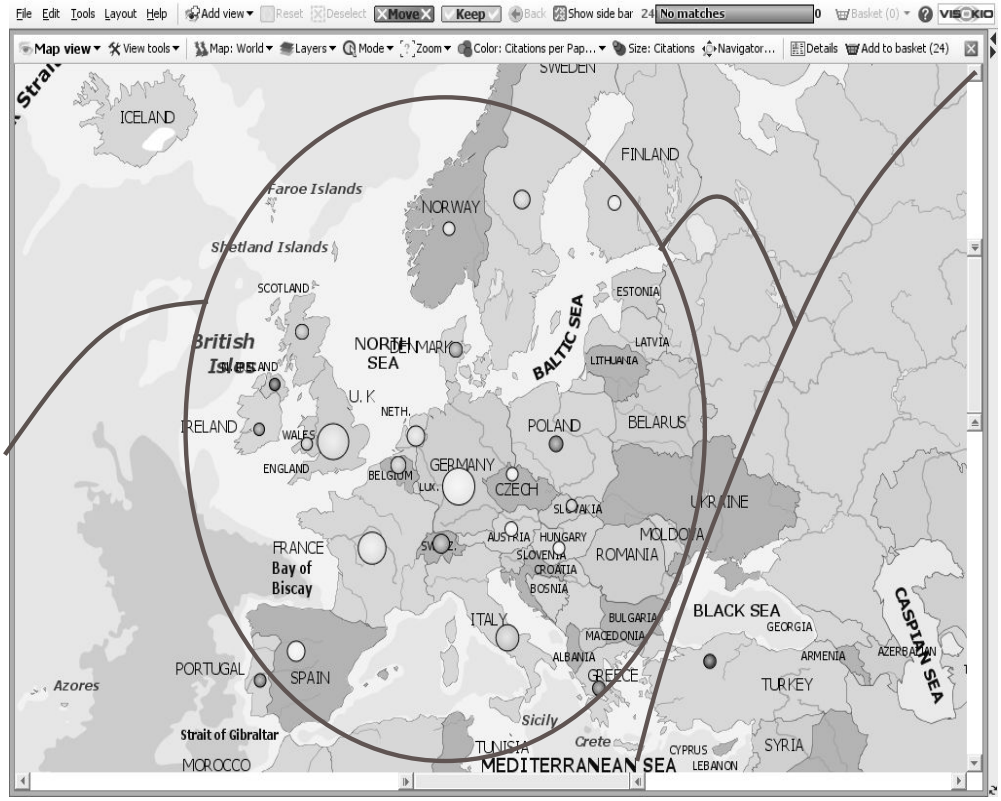

Figure 23. Europe and **Turkey** on the Productivity Belt (Engineering Field; sizes of glyphs show Citations)

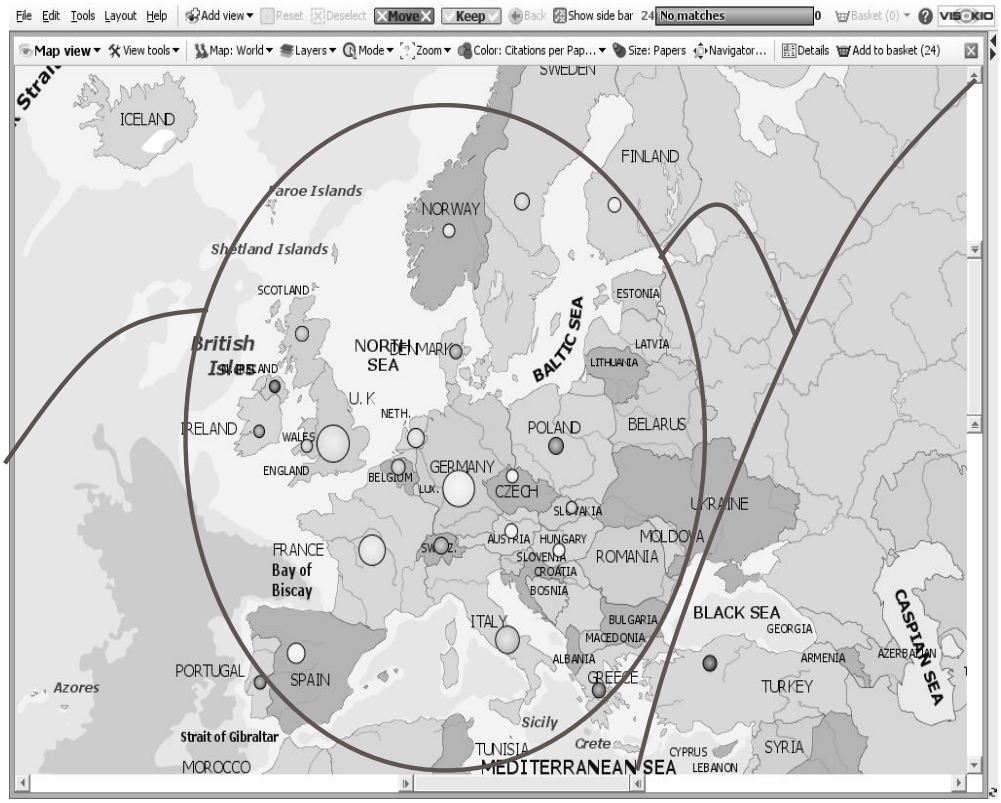

Figure 24. Europe and **Turkey** on the Productivity Belt (Engineering Field; sizes of the glyphs show Papers)

We have also linked the demographic data to our analyses for each field. We have taken the Population values of each Country and then computed the following ratio for all three fields:

$$\text{PPTP} = (\text{Papers}/\text{Population})*1000,$$

which is the mathematical formulation of Papers Per Thousand Population. Figure 25 denotes the PPTP distribution of the countries in Engineering Field. More than %60 of the countries have a PPTP value less than 0.2 and nearly half of them have a PPTP value less than 0.1. In Figure 26, we observe even more countries having a PPTP value less than 0.2, precisely %63 of the countries publishing in Computer Science. In E&B Field, this same percentage reaches up to %80 of the countries (Figure 27). If we knew or at least had a good guess on the fraction of Population that carry out research, we could have divided the PPTP value with this fraction to find out the number of papers per researcher in various fields for each country and this would be another important metric to consider.

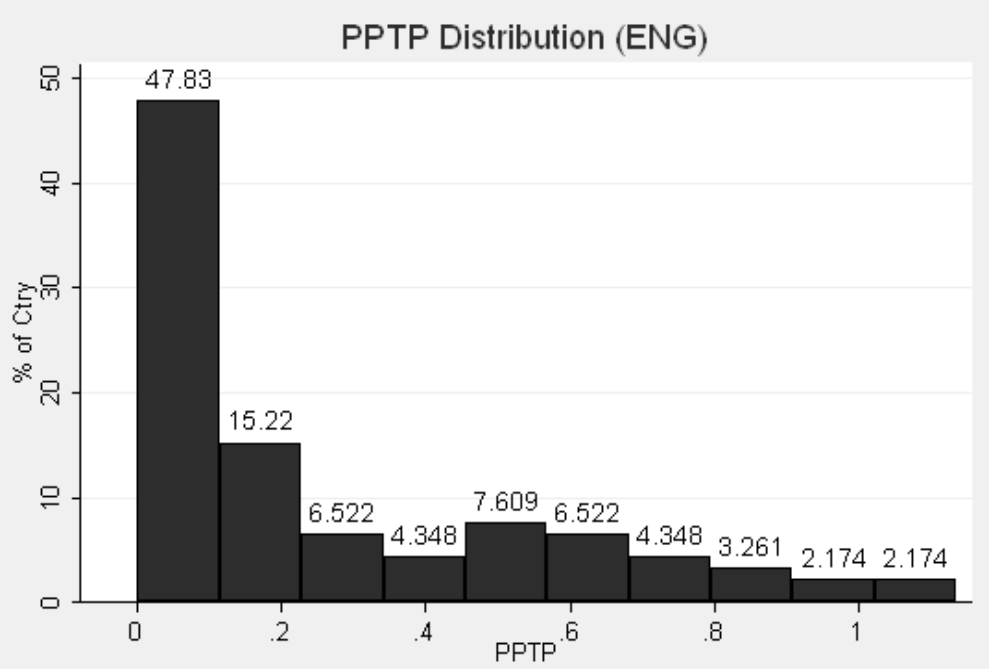

Figure 25. PPTP Distribution in Engineering Field (**Singapore** excluded)

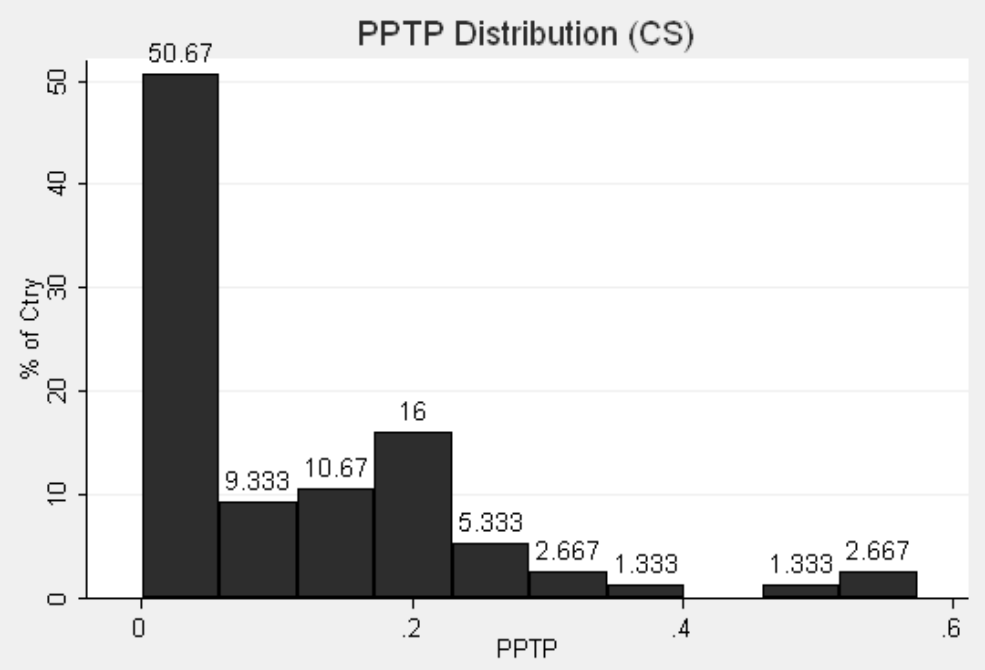

Figure 26. PPTP Distribution in Computer Science Field

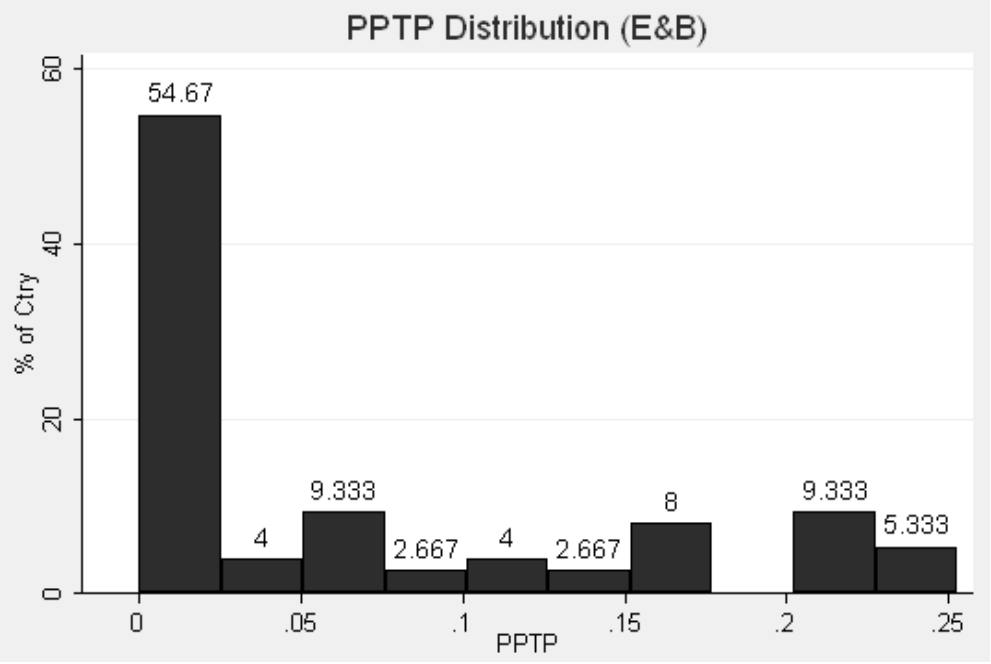

Figure 27. PPTP Distribution in Economics & Business Field

Another analysis that we carried out was to plot the PPTP and CPP values of the countries against each other. To do this, we have used Miner3D once more and plotted the data for each Field. The x-axis, marker size and elevation show the PPTP value while the y-axis shows the CPP value. The color indicates the Continent of the Country.

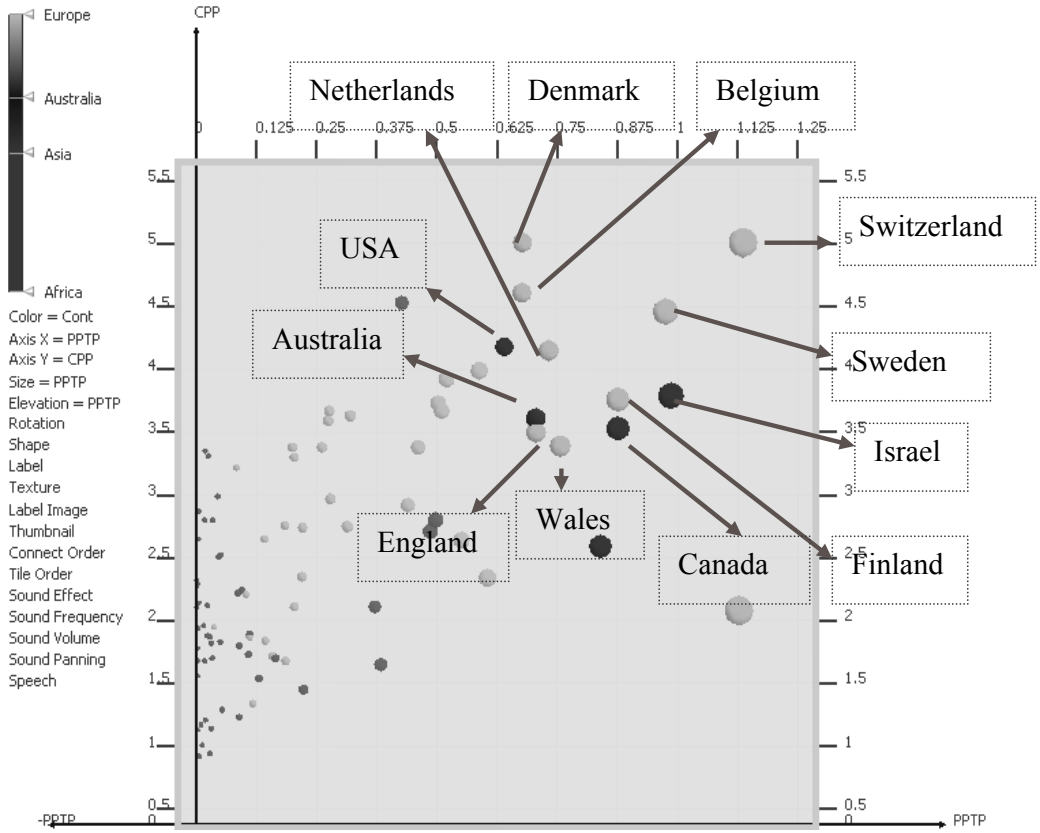

Figure 28. CPP-PPTP Plot of Engineering (Continent focused, **Monaco** and **Singapore** excluded)

As seen in Figure 28, the top-right corner of this plot is mostly dominated with a green color (which denotes Europe) and these European countries are **Switzerland, Sweden, Belgium, Denmark, Netherlands** and **Finland**. There are also two more countries from America (**USA** and **Canada**) and one country from Asia (**Israel**). All of these countries demonstrate both highly influential academic output (high CPP values) and scientific productivity as a nation in Engineering Field when PPTP is considered.

In Computer Science (Figure 29), we observe a different pattern. In the top-right quarter of the plot, we only have a single Country, namely **Israel.** Apart from her, we observe two more countries on the bottom-right corner (meaning a low CPP and a high PPTP value), namely **Singapore** and **South Korea** and two more countries on the top-left corner (meaning a high CPP and a low PPTP value), namely **USA** and **Hong Kong.** From the remaining countries, **Sweden, Denmark** and **Switzerland** again appear just on the top-right corner.

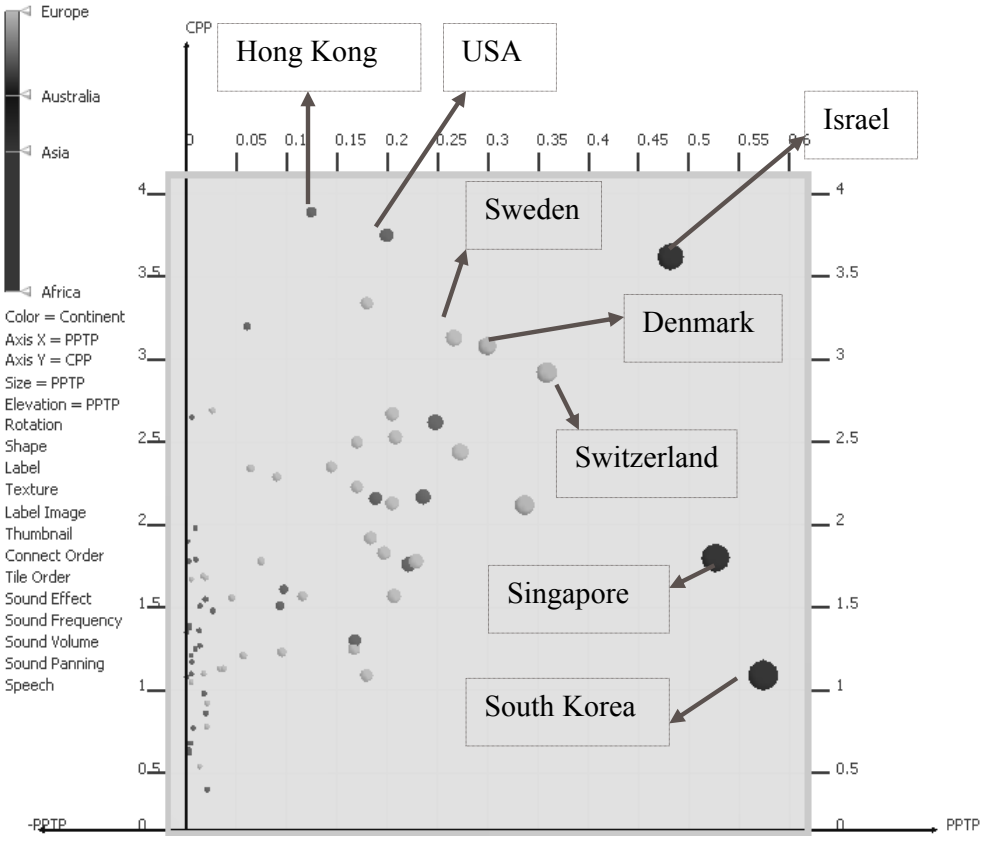

Figure 29. CPP-PPTP Plot of Computer Science (Continent focused, **Liechtenstein** excluded)

Finally in Economics & Business (Figure 30), we again observe only two countries in the top-right corner, namely **USA** and **Israel**. **Hong Kong** appears at the top according to CPP but in the middle according to PPTP. The remaining countries in this Field seem to be positioned at the bottom of the diagonal, meaning that they have a better position with respect to x-axis (PPTP axis) than y-axis (CPP axis). Some of the remarkable countries on the plot are **Sweden, Canada, England** and **Netherlands** in E&B Field.

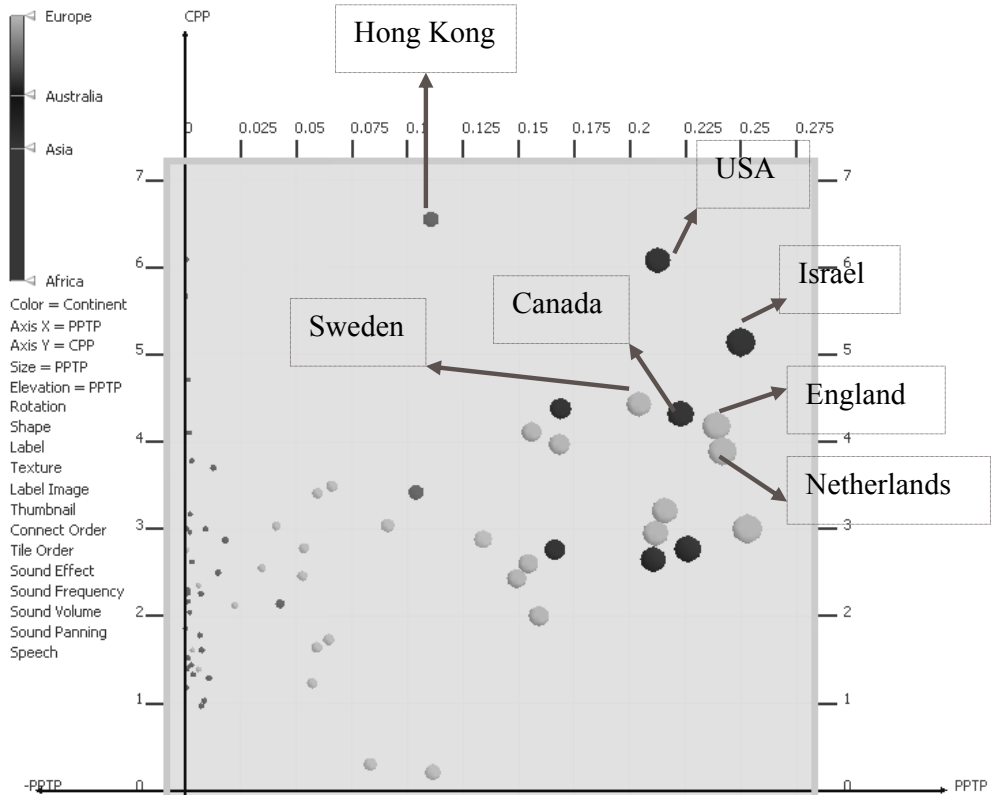

Figure 30. CPP-PPTP Plot of Economics & Business (Continent focused, **Macao & Ecuador** excluded)

The last analysis performed in this study is the simultaneous analysis of Population and PPTP values of the countries in all three fields. In Figures 31-36, we first demonstrate the distribution of PPTP and Population by using color and size of the tiles respectively. For each Field, we then add a view where the tiles are grouped by Continent. By doing this, we aim at understanding which continents in the world are academically productive and the result is very interesting.

In Engineering (Figure 31 and 32), we observe green colors in American, European, Australian and Asian countries with Europe having the largest number of productive nations. However, the general picture of Engineering seems to be dramatic as can be read from the dominant dark red colors.

In Computer Science (Figure 33 and 34), the picture is much worse than Engineering. We hardly ever observe green colors, some in Asia and some in Europe.

In Economics & Business (Figure 35 and 36), we again have a similar picture to the one in Engineering. America, Europe, Australia and Asia have a few nations which are academically productive but the rest is totally unproductive.

Figure 31. PPTP-Population Relation in Engineering Field (**Singapore** excluded)

Figure 32. PPTP-Population Relation in Engineering Field (Continent focused, **Singapore** excluded)

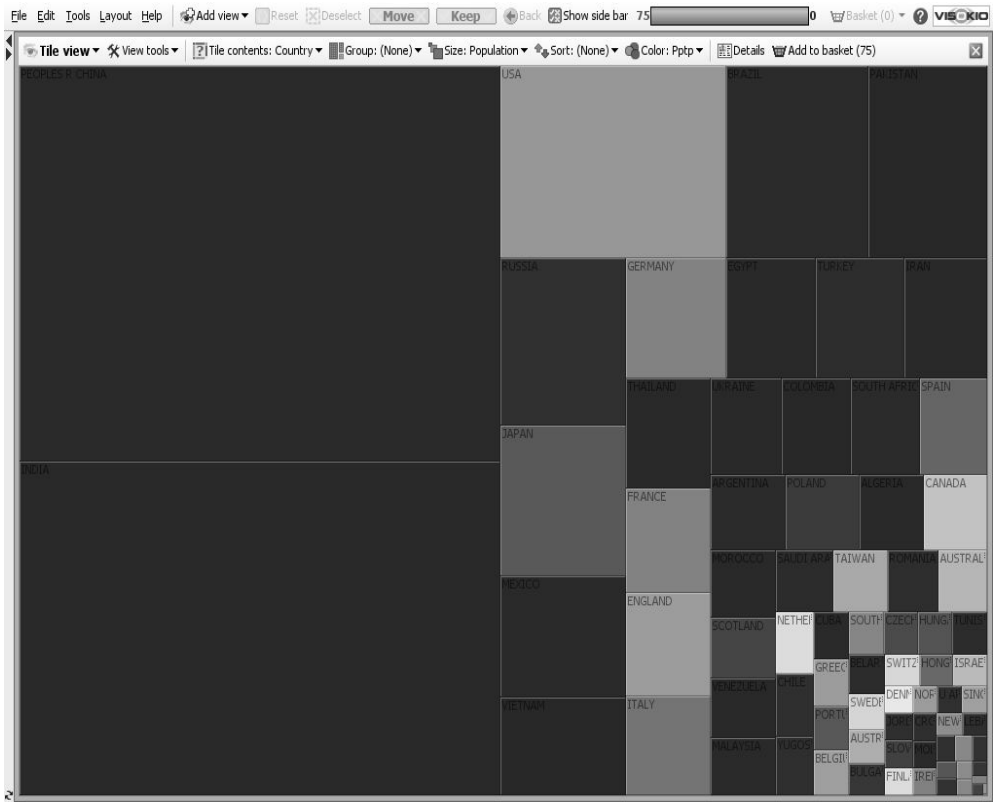

Figure 33. PPTP-Population Relation in Computer Science Field

Figure 34. PPTP-Population Relation in Computer Science Field (Continent focused)

Figure 35. PPTP-Population Relation in Economics & Business Field

Figure 36. PPTP-Population Relation in Economics & Business Field (Continent focused)

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed application of visualization schemes for benchmarking the publication performance of world countries in selected fields of science. We deployed software tools to visualize evaluate the main performance metrics used in academia; namely Papers, Citations and Citations per Paper (CPP). Meanwhile, we have included other dimensions in our study, such as Population and Papers per Thousand People (PPTP) with the goal of gaining further insights. Some of our analyses involved mapping the data to geographical coordinates, as well.

Our study can be extended in several ways:

1) Our analyses can be extended to cover all 22 fields of science. Such an analysis would serve as a reference source for further studies.

2) Other country statistics can be included in the analysis as data dimensions. Some candidate dimensions are

    i. number of patents for each Country for each Field (hence patents/paper metric can be introduced)
    ii. GNP of each Country (expressed in terms of same currency)
    iii. number of researchers in research and development

   All the three indicators mentioned above can be found in WDI Online Database (WDI Online) from which we acquired the major part of the population data. WDI also provides data of the number of scientific and technological journal articles throughout years.

3) The data can be analyzed through Data Envelopment Analysis (DEA) (Cooper et al., 2006). DEA is one technique based on optimization for benchmarking a given set of Decision Making Units (DMUs) with multiple inputs and outputs. In DEA, inputs are the resources that the DMUs utilize (such as workforce, time, funds) and the outputs are the products or services that DMUs produce. For example, in our context, one could specify the countries as DMUs, Population, GNP, Number of Researchers as inputs and Papers and Citations as outputs. In DEA, the input and output data are processed and the *efficiency* of each DMU is expressed as a single rational number between 0 and 1, named as the *efficiency score*. The DMUs which have efficiency scores of 1 are considered *efficient*, and the other DMUs are considered *inefficient*. The reference set suggests which of the efficient DMUs the inefficient DMU should take as reference in order to increase its efficiency. We believe that carrying out DEA is the most fruitful path of research to extend our study.

4) One can use formal statistical techniques in analyzing the data and for drawing strongly supported conclusions. For example, one can formally test the hypothesis of whether the CPP values for different fields are statistically not the same by using single factor ANOVA (Rice, 1995). Furthermore, the remarks on possible distributions to fit (based on the

visual inspection of histograms) can be turned into solid statements through goodness-of-fit tests.

5) One can use well-known data mining techniques to obtain further insights. For example, one can carry our association mining to find out rules such as "In %X of countries, having CPP > 3 in field A implies having CPP > 3 in field B"

We hope that our study will serve as a useful example of how visualization schemes in information visualization field can be used in technology management. We conclude by remarking that our study reveals specific answers to only certain types of questions.

REFERENCES

Ashton, W. Bradford; Sen, Rajat K. (1989) Using Patent Information in Technology usiness Planning-II. *Research Technology Management;* Jan/Feb 1989; 32, 1; ABI/INFORM Global pg. 36

Boyack, K. W., Wylie, B. N., Davidson, G. S. (2002) Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology;* JUl 2002; 53, 9; ABI/INFORM Global pg. 764

Butz, A., Schmidt, A. (2005) Vorlesung Advanced Topics in HCI (Mensch-Maschine-Interaktion 2). Available under http://www.medien.ifi.lmu.de/fileadmin/mimuc/mmi2_ss05/vorlesung/2005-06-08_f.pdf

CIA. https://www.cia.gov/cia/publications/factbook/

Cooper, W. W., Seiford, L. M., & Tone, K. 2006. *Introduction to data envelopment analysis and its uses***.** New York, NY: Springer.

Erickson, S. G. (1996) Using patents to benchmark technological standing: international differences in citation patterns. *Benchmarking for Quality Management & Technology*, Vol: 3 No: 1, 1996, pp. 5-18. MCB University Press, 1351-3036

Ertek, G. and Demiriz, A. (2006). A framework for visualizing association mining results. *Lecture Notes in Computer Science*, Vol: 4263, pp. 593-602. Available under http://www.ertek.info

Ertek, G., Incel, B. K., Impact of Cross Aisles in a Rectangular Warehouse: A Computational Study, INFORMS International 2006, June 2006, Hong Kong, China. Available under http://www.ertek.info

ESI Database. http://scientific.thomson.com/products/esi/

GapMinder. http://www.gapminder.org/

Ingwersen, P., Larsen, B. (2001) Mapping National Research Profiles in Social Science Disciplines. *Journal of Documentation*, Vol: 57, No: 6, November 2001, pp. 715–740

Java. http://java.sun.com

Leta, J. (2005) Human resources and scientific output in Brazilian science: Mapping astronomy, immunology and oceanography. *Aslib Proceedings: New Information Perspectives*, Vol: 57 No: 3, 2005 pp. 217-231

Many Eyes. http://services.alphaworks.ibm.com/manyeyes/home

Miner 3D Enterprise. http://www.miner3d.com/products/enterprise.html

Mondrian. http://rosuda.org/Mondrian/

NationMaster. http://www.nationmaster.com

OECD. http://www.oecd.org/home/0,2987,en_2649_201185_1_1_1_1_1,00.html

Omniscope. http://www.visokio.com/omniscope/features.cfm

Plaisant, C. (2004) *The challenge of information visualization evaluation*. ACM Press, New York, NY, USA.

Rice, J. A. (1995) *Mathematical statistics and data analysis.* Thomson Publishing.

RoSuDa. http://stats.math.uni-augsburg.de/software/

SK Card, JD Mackinlay, B Shneiderman (1999) *Readings in information visualization: using vision to think.* Morgan Kaufmann.

Stata. http://www.ats.ucla.edu/stat/stata/faq/odbc.htm

Swivel. http://www.swivel.com/

Thesaurus. http://thesaurus.reference.com/

Thomson Scientific. http://scientific.thomson.com/products/solutions/acad/

Tufte, E. (2001) *The visual display of quantitative information.* Graphics Press, CT, USA.

Ulus, F., Kose, O., Ertek, G. Sahin, G. (2006) Visualizing the Data Envelopment Analysis. Available under http://www.ertek.info

UN. http://www.un.org/english/

Visokio. http://www.visokio.com

WDI Online.
http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20398986~menuPK:64133163~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html

Wikipedia. http://www.wikipedia.org/