Replication of working paper: Migration and Innovation: Learning from Patent and Inventor Data

Name: Manvanth Sunadreshan

Email: Sudareshan.manvanth@stud.hs-fresenius.de

ID: 400353080

Last compiled on 22 July, 2024

Table of Contents

1 Abstract

This working paper, titled “Migration and Innovation: Migration and Innovation: A Survey of Recent Empirical Research” by Francesco Lissoni & Ernest Miguelez applies methodologically rigorous patent and inventor data to the study of migration and innovation’s complex connections.

Based on archival data, the work investigates what knowledge transfer mechanisms are used, how migration effects inventor activity in innovation systems of the destination country, and what specific impacts migrant inventors produce. It gives information about data pattern and movement and also dissects the functions of both the top and novice migrant inventors in distributing knowledge across the globe.

By identifying the two fields of economic geography and migration economics and integrating empirical evidence from these two literatures, this paper demonstrates that the essence of these fields is in knowledge transfer between people. The final section covers strengths as well as limitations of the different approaches used to define migrant inventors including, nationality coverage in patents, name matching and merging of patents with administrative databases.

2 Introduction

In this working paper Migration and Innovation: A Survey of Recent Empirical Research, the authors Francesco Lissoni & Ernest Miguelez explore some nuances in that relationship using comprehensive patent/inventor data. The report has a broad remit to analyse the contributions of migrant inventors by going deeper into understanding how they fit in with innovation ecosystems at their host countries – what mechanisms work best for transferring knowledge, dynamics behind inventor mobility and on local level impact.

The content is designed in a well-planned way, with the basic discussion of how to access and process patent data at start followed by what type information can be extracted on migrant inventors. Then it reports descriptive statistics on main data trends and problems. It also examines the experiences of both senior and junior migrant inventors, demonstrating their role in global knowledge dissemination with some finer points within this group.

Analyzing empirical research from economic geography and thus migration economics, these subject areas are shown to intersect, with the former focusing on the analysis of interpersonal knowledge transfer. In the conclusion of the paper, the author examines the strengths and weaknesses of the ways of identifying such migrant inventors as nationality of coverage in patents; name identification and the advantages and drawbacks of matching patent information with the administrative data. It is necessary for studying an effect of the global dynamics of inventor migration and its impact on the innovation networks and technological developments.

3 Objective

The objective of this paper is to comprehensively examine different facets of migrant inventors' impact on innovation of the host countries. By replicating this study, it is the hope to employ the measure to assess the contribution of such inventors on knowledge transfer, the reduction of the R&D labor shortage and the boosting of the diversity of inventive teams. Since the research intends to investigate trends and impact of the international mobility of inventors from the patent information analysis it will endeavor to produce empirical evidence of how migration induces technological advancements and economic development. Also, the paper aims to extend the understanding of heterogeneity of migrant inventors from highly qualified stock holders of knowledge to junior innovators and their contribution to the maintenance of innovation, let alone the stimulation of innovation systems.

4 Methodology

Before diving into the analysis of strategies employed into the replication strategy, we shall first understand R and its functionalities and we shall also go through the data collection process.

4.1 R

R is a powerful, flexible, multifunctional computer language and application that has been developed as an open-source software for the purpose of statistics and graphical procedures. Ranges of functions of R, are much more extensive than of Python: linear and nonlinear regression, traditional statistical tests, time series analysis, classification, and clustering; it was developed by statisticians Ross Ihaka and Robert Gentleman. Another one is comprehensive with a library containing a large number of packages which could be downloaded from the Comprehensive R Archive Network. Being an open-source language with efficient features in handling data and excellent data visualization accompanied by a strong community back up, R is commonly used in academic and business world for data analysis and machine learning.

4.2 R Markdown

R Markdown is a versatile authoring system that combines the R code with the Markdown to produce documents and interactive reports, presentations, and dashboards. It enables a user to insert R codes inside the Markdown text; it then produces documents that include both the code and its output like a graph or a table. Being compatible with HTML, PDF, Word, and other output formats, R Markdown is suitable for using in reproducible research, data analysis, and documentation. Therefore, it complements and prepares a more integrated narrative combined with code and results in one document that increases interactivity and forms of reports and analyses.

4.3Data Collection

The data collection and analysis for the replication package “Migration and Innovation: Step by step, ‘Analyzing Patent and Inventor Data’ comprises of the following steps. For the first data source, USPTO PatentsView is employed because it contains all the granted patents’ details including inventors’ and assignees’ names reported as of 16th of October, 2019. It is also used to generate the unique inventor identifiers and patents grouped according to technological categories. Furthermore, the study uses information on IBM’s Global Name Recognition to identify the nationality of inventors according to their names. Moreover, the data contain information from PCT patents which is useful for international analysis of patenting phenomena. The replication package uses Stata to handle data and conduct analysis and all the necessary scripts are included in the master do-file to produce the figures in the study. The raw and processed data collected in the study are presented as a single dataset to provide the reader with an idea of the data available for analysis and to enhance the study’s replicability.
The “uspto_inventors_coords. dta” dataset is also recognized as a part of this replication package. Specific ideas for this were obtained from the USPTO PatentsView database and involve inventors’ ID and their associated country of origin, the year the patent was granted, and the primary technological class of the granted patent.
For the replication of the Five graghical representaions in this study, different data sources are given in the readme file,they are:For the replication of the Five graghical representaions in this study, different data sources are given in the readme file,they are:

1. Figure 1: Uses “. \Data\Created-data\uspto_inventors_coords. dta” to present the results concerning the number of patents by the inventor.

2. Figure 2: Also uses “. \Data\Created-data\uspto_inventors_coords. dta” as a label to indicate the percentage distribution of the inventors according to the country they reside in.

3. Figure 3: Integrate the USPTO data used in “. \Data\Created-data\uspto_inventors_coords. dta” with the IBM-GNR processed data in “. \Data\Created-data\inventor_gnr_strict. dta” to exhibit the nationality of the inventors with the help of both PCT and USPTO databases.

4. Figure 4 repeats the study of Figure 3 across the technological fields by using other datasets, such as “. \Data\Raw-data\WIPO\Bilateral flows Electrical engineering sector. dta” and others for various technological fields.

5. Figure 5: Attempts to replicate what is done in figure 4, calculating the forward and backward citations of US-based inventors and using “. \Data\Created-

data\uspto_inventors_coords. dta", ". \Data\Created-data\inventor_gnr_strict. dta", and citation data from ". \Data\Raw-data\g_us_patent_citation. tsv".

4.4 Replication process

I commenced the process of replication through R by selection of the working paper in the pool of working papers provided by our professor. I then completed downloading the paper's replication package equaling the size of approximately 24 Gigabits. After careful examination of the replication package, I was able to identify it contained various files such as Read me file, Code file, data file, output file, 2 stata files related to the working paper.

I was able to identify the data used to replicate each graph through the help of read me file. I immediately launched R studio software on my computer system and installed all the necessary packages required to support my replication process. As I was provided with stata file I was able to invoke the necessary data with ease.

`ggplot2` is a powerful and flexible R package for data visualization. It is part of the `tidyverse` collection and is known for its ability to create complex and aesthetically pleasing visualizations with concise, readable code. The package uses a grammar of graphics approach, allowing users to build plots layer by layer. This approach makes it easy to customize and extend plots, helping users to effectively communicate their data insights. The use of `ggplot2` in this script ensures that the resulting plot is clear, informative, and visually appealing, which is crucial for accurately conveying the distribution of patents per inventor.

For data visualization ggplot2 package must be installed in R. It is a part of tidy verse package and is able to generate complex and intricate visualizations with condensed and readable code. Through the use of ggplot2 reader can build data visualizations layer by layer. The tidyverse allows the replicator to manipulate data as per requirement with ease and for easy analysis and can be installed using install.packages("tidyverse").

The installation of necessary packages and the invocation of packages was conducted using if else statement to avoid multiple installation of the already installed packages. The working directory can be done using the function setwd() or manually by going to session and changing directory to working directory.

```
if (!require(tidyverse)) {
  install.packages("tidyverse")
  library(tidyverse)
} else {
  library(tidyverse)
}

if (!require(haven)) {
  install.packages("haven")
  library(haven)
} else {
  library(haven)
```

```
}

if (!require(dplyr)) {
  install.packages("dplyr")
  library(dplyr)
} else {
  library(dplyr)
}
```

4.5 Issues faced

I chose to do it manually as my files does not contain "_" in between words for each file name and it is not functioning as expected and displaying error.

The following code I have imported the data for replicating initial plot from stata file defining the path of the file. The view() function is used to view the data set.

```
uspto_inventors_coords <- read_dta("D:/Hochschule Fresenius notes (sem2)/SS
2024- Data Science in business/Final_Project/Data/Created-
data/uspto_inventors_coords.dta") #file path

View(uspto_inventors_coords)
```

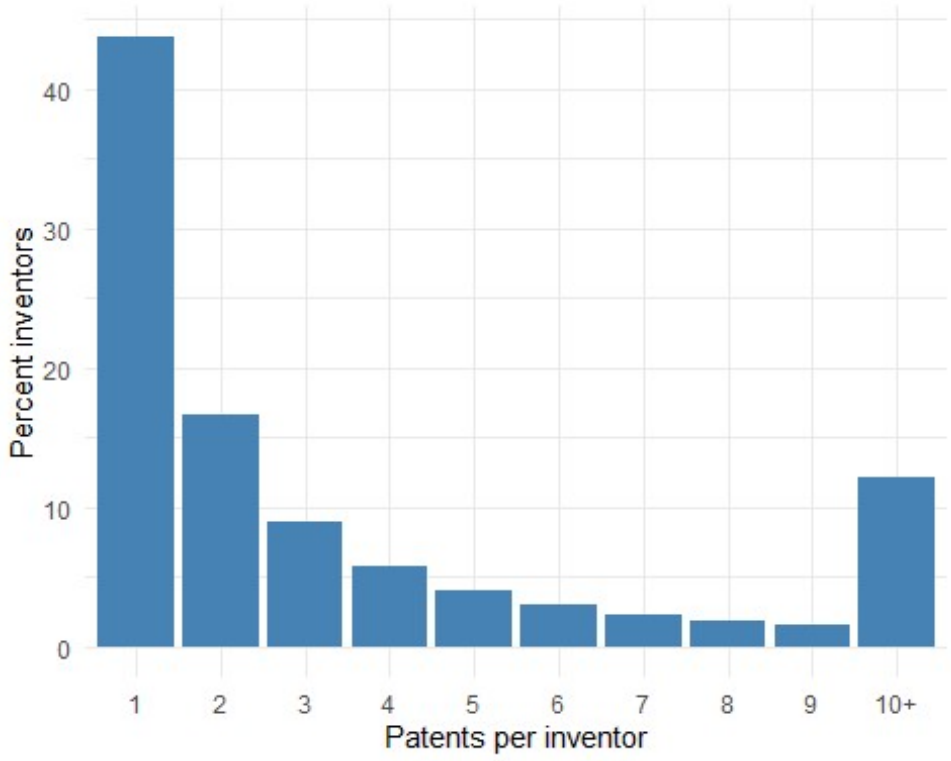
5 Graphical Representation in "R" for Share of Inventors

```
inventor_patents <- uspto_inventors_coords |>
  group_by(inventor_id) |>
  summarise(patent_count = n())

patent_distribution <- inventor_patents |>
  count(patent_count) |>
  mutate(percent_inventors = n / sum(n) * 100)

patent_distribution <- patent_distribution |>
  mutate(patent_count = ifelse(patent_count >= 10, "10+",
as.character(patent_count)),
         patent_count = factor(patent_count, levels = c(as.character(1:9),
"10+")))

ggplot(patent_distribution, aes(x = patent_count, y = percent_inventors)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Patents per inventor", y = "Percent inventors") +
  theme_minimal()
```



To create count the number of patents per inventor this code is groups the dataset by inventor_id and calculates the number of patents for each inventor, creating a new data frame inventor_patents with the inventor IDs and their respective patent counts. To create a new data frame with the percentage of inventors per patent counts the occurrences of each patent_count value and calculates the percentage of inventors for each count, resulting in a data frame patent_distribution that includes the patent count and the corresponding percentage of inventors.

We then Bin the patent into categories, grouping all counts of 10 or more into a "10+" category. It then ensures that the patent_count variable is treated as a factor with specified levels from 1 to 9 and "10+". AT the end we use the function of ggplot() to generate the Visualization of figure 1.

By implementing the above code we can replicate the figure identical to the plot given in the working paper.

6 Graphical Representation in "R" for Distribution of Inventors' Country of Residence

```
uspto_inventors_coords <- tibble(
  country = c('US', 'Europe', 'Japan', 'Korea', 'China', 'Others'),
  `1976-1990` = c(50, 20, 15, 5, 1, 9),
  `2006-2020` = c(40, 25, 10, 8, 7, 10)
)
```

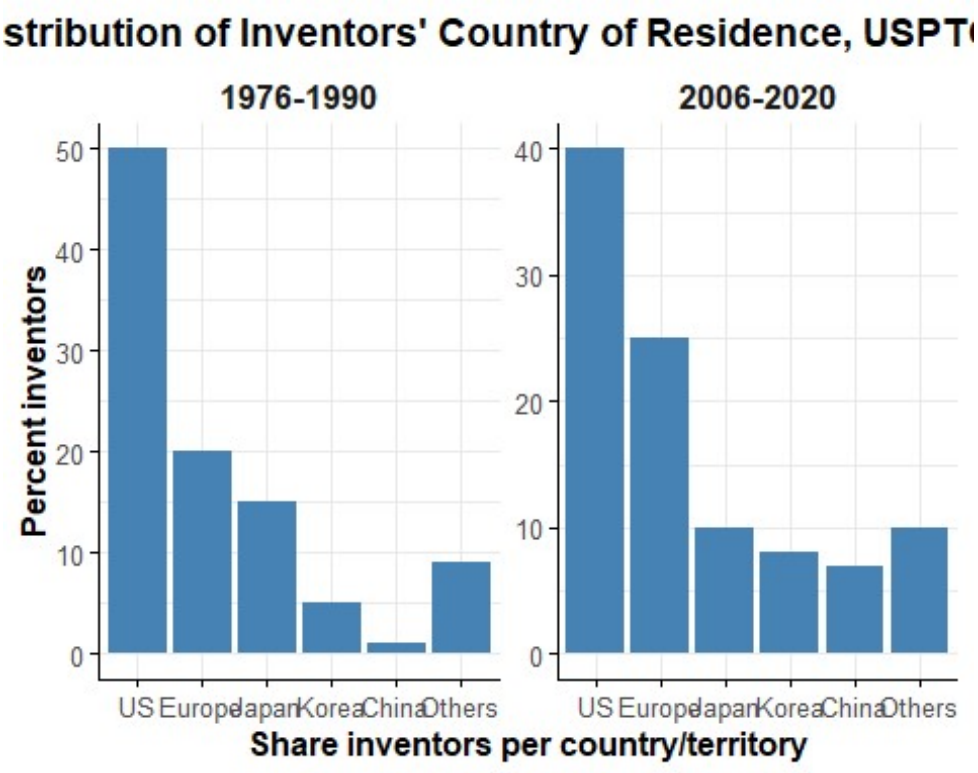
```
uspto_inventors_coords_long <- uspto_inventors_coords |>
  pivot_longer(cols = `1976-1990`:`2006-2020`, names_to = "period", values_to
= "percent_inventors")

uspto_inventors_coords_long <- uspto_inventors_coords_long |>
  mutate(country = factor(country, levels = c('US', 'Europe', 'Japan',
'Korea', 'China', 'Others')))


uspto_inventors_coords_long <- uspto_inventors_coords |>
  pivot_longer(cols = `1976-1990`:`2006-2020`, names_to = "period", values_to
= "percent_inventors")

uspto_inventors_coords_long <- uspto_inventors_coords_long |>
  mutate(country = factor(country, levels = c('US', 'Europe', 'Japan',
'Korea', 'China', 'Others')))

ggplot(uspto_inventors_coords_long, aes(x = country, y = percent_inventors,
fill = country)) +
  geom_bar(stat = "identity", fill = "steelblue", show.legend = FALSE) +
  facet_wrap(~period, scales = "free_y", ncol = 2) +
  labs(title = "Distribution of Inventors' Country of Residence, USPTO Data",
       x = "Share inventors per country/territory",
       y = "Percent inventors") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    strip.text = element_text(size = 13, face = "bold"),
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(size = 10, angle = 0, hjust = 0.5),
    axis.text.y = element_text(size = 10),
    axis.line.x = element_line(color = "black"),
    axis.line.y = element_line(color = "black"),
    axis.ticks = element_line(color = "black")
  )
```



By observing figure 2, we can observe there are dataset of investors' country of residence fro 2 periods, 1976-1990 and 2006-2020, and we must visualize this data using a bar plot.To generate this graph I have used a function facet_wrap which allows me to plot both graphs.

6.1 Issues faced

As we can clearly see the plot generated does not align with the plot given in the working paper. The attempt to trouble shoot has not been successful and I was unable to identify the fault in my code.

Further guidance from my professor will definitely aid me in identifying the problem in my code and due to various time constraints I was not able to successfully contact my lecturer in time to complete the replication of this graph.

7 Various facets of the replication process

7.1 Tools used

Th study supporting my replication process is supported by the lecture notes provided by Prof. Dr. Stephan Huber, which can be accessed at [https://hubchev.github.io/ds/] .

7.2 Additional tools used during replication

1. Online tools: Chatgpt, statistica,other visualization solutions and several internet-based tools.

2. R studio software.

7.3 Issues faced with replication

The experience of learning "R" language was definitely new and challenging as I have no prior experience in similar software. My require more exposure to situations that require troubleshooting in order for me to better understand the functionality of R.

Though the data was conveniently given in required format, the size of the dataset was too large for the system to handle. It took me sufficient amount of time to read in a single dataset and even more time to generate the output after compilation.

7.4 Future scope

As I was not able to grasp of the concepts that were introduced to me in R, it has constrained my efficiency to replicate the working paper more effectively and efficiently. If I possessed sufficient in depth knowledge on R I would have understood the different ways to solve the issues. And also I would conduct an extensive analysis and would be able to visualize data in a far for comprehensible manner.

8 Conclusion

Technology and science, innovation processes also have always been related to migration, since migrants made essential contributions to these fields. This contribution is clear in the historical movement and the current mobility of highly qualified people. This trend is due to the fact that migrant inventors complement R & D in labor markets where they have deemed there is a shortage of diversity in skills relating to innovation.

No less important, patent and inventor data present a picture contrary to the alleged marginality and inefficiency of migrants: on the contrary, migrants are business-minded, hardworking, creative, and determined in their attempts to contribute to the advancement of knowledge across borders. Thus, the mutual connection between migration and innovation points out the need to promote skills mobility for the sake of technological advancement.

Through this project, I was able to understand the various gaps in the knowledge I have regarding the use of R. I was able to experience the world of Git and GitHub. In summary this software has a great potential for the future.

9 Contribution using Git and Github

Git is a type of version control that operates in a distribute platform whereby many developers can operate in different branches but still work on the same project. This increases the flexibility of the user to navigate through the project with ease.

GitHub is a web-based tool that incorporates Git as its base but also enlarges it with the more graphic user interface, hosting services for Git repositories plus some added-blow tools like issue tracking, project management and collaboration tools. Collectively they enable optimal working, and versioning of software programs.

In this project one of the requirements was to make a Github contribution. We shall see a short example on how github contribution,

First I created a Github profile by uploading all the necessary information.

I also installed git onto my system through my browser. After installation, I commenced the process on Git push.

git push:

'git push' is a Git command used to upload local repository content to a remote repository.When you make changes to your local repository and commit them, these changes are stored locally.

To share these changes with others or to backup your work on a remote server (like GitHub), you use git push.It sends the committed changes from your local repository to the corresponding branch in the remote repository.

Its Syntax: git push

```
git push origin main
```

In this example, origin is the default name of the remote repository, and main is the branch to which the changes will be pushed.

The files for the working project can be found on my repository in Github which can be accessed by using []

We also made a pull request, the tutorial can be that can be accessed through [https://github.com/hubchev/make_a_pull_request]. We must go through the readme file and understand the complete process of pull request.

10 Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and

authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I have read the Handbook of Academic Writing by Hildebrandt & Nelke (2019) and have endeavored to comply with the guidelines and standards set forth therein.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

The report includes:

- ☒ About 4000 words (+/- 500).
- ☒ A title page with personal details (name, email, matriculation number).
- ☒ An abstract.
- ☒ A bibliography, created using BibTeX with APA citation style.
- ☒ The complete R code required to reproduce the results.
- ☒ Detailed instructions on data acquisition and importation into
- ☒ An introduction to guide the reader and a conclusion summarizing the work and discussing potential future extensions.
- ☒ All significant resources used in the report and R code development.
- ☒ The filled out Affidavit.
- ☒ A concise description of the successful use of Git and GitHub, as detailed here: - - make_a_pull_request.
- ☒ A concise description of the presentation published on GitHub.
- ☒ The project submission includes:
- ☒ … The .qmd file(s) of the report.
- ☒ … The _quarto.yml file of the report.
- ☒ … The .pdf file of the report.
- ☒ … The standalone .html file of the report.
- ☒ … All necessary files (not available online) to reproduce the report and the R code.

- ☒ … The standalone .html file of the presentation.

ManvanthSundareshan

22.07.2024

Cologne, *Germany*

11 References

Lissoni, Francesco, and Ernest Miguelez. 2024. “Migration and Innovation: Learning from Patent and Inventor Data.” Journal of Economic Perspectives, 38 (1): 27–54.