



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2018/10.17.00.07-TDI

A DATA SCIENCE APPROACH TO LATTES CV DATA ANALYSIS

Thiago Luís Viana de Santana

Master's Dissertation of the
Graduate Course in Applied
Computing, guided by Dr.
Rafael Duarte Coelho dos Santos,
approved in September 20, 2018.

URL of the original document:

<http://urlib.net/8JMKD3MGP3W34R/3S3AQHH>

INPE
São José dos Campos
2018

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GBDIR)

Serviço de Informação e Documentação (SESID)

CEP 12.227-010

São José dos Campos - SP - Brasil

Tel.:(012) 3208-6923/7348

E-mail: pubtc@inpe.br

**COMMISSION OF BOARD OF PUBLISHING AND PRESERVATION
OF INPE INTELLECTUAL PRODUCTION (DE/DIR-544):****Chairperson:**

Dr. Marley Cavalcante de Lima Moscati - Centro de Previsão de Tempo e Estudos Climáticos (CGCPT)

Members:

Dra. Carina Barros Mello - Coordenação de Laboratórios Associados (COCTE)

Dr. Alisson Dal Lago - Coordenação-Geral de Ciências Espaciais e Atmosféricas (CGCEA)

Dr. Evandro Albiach Branco - Centro de Ciência do Sistema Terrestre (COCST)

Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia e Tecnologia Espacial (CGETE)

Dr. Hermann Johann Heinrich Kux - Coordenação-Geral de Observação da Terra (CGOBT)

Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação - (CPG)

Silvia Castro Marcelino - Serviço de Informação e Documentação (SESID)

DIGITAL LIBRARY:

Dr. Gerald Jean Francis Banon

Clayton Martins Pereira - Serviço de Informação e Documentação (SESID)

DOCUMENT REVIEW:

Simone Angélica Del Ducca Barbedo - Serviço de Informação e Documentação (SESID)

André Luis Dias Fernandes - Serviço de Informação e Documentação (SESID)

ELECTRONIC EDITING:

Marcelo de Castro Pazos - Serviço de Informação e Documentação (SESID)

Murilo Luiz Silva Gino - Serviço de Informação e Documentação (SESID)



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2018/10.17.00.07-TDI

A DATA SCIENCE APPROACH TO LATTES CV DATA ANALYSIS

Thiago Luís Viana de Santana

Master's Dissertation of the
Graduate Course in Applied
Computing, guided by Dr.
Rafael Duarte Coelho dos Santos,
approved in September 20, 2018.

URL of the original document:

<<http://urlib.net/8JMKD3MGP3W34R/3S3AQHH>>

INPE
São José dos Campos
2018

Cataloging in Publication Data

Santana, Thiago Luís Viana de.

Sa59d A data science approach to Lattes CV data analysis / Thiago Luís Viana de Santana. – São José dos Campos : INPE, 2018.
xxii + 76 p. ; (sid.inpe.br/mtc-m21c/2018/10.17.00.07-TDI)

Dissertation (Master in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2018.

Guiding : Dr. Rafael Duarte Coelho dos Santos.

1. Data analysis. 2. Data science. 3. Lattes platform.
4. Bibliometrics. 5. Artificial intelligence. I.Title.

CDU 001.103:004.8



Esta obra foi licenciada sob uma Licença [Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada](https://creativecommons.org/licenses/by-nc/3.0/).

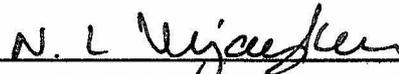
This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](https://creativecommons.org/licenses/by-nc/3.0/).

Aluno (a): **Thiago Luís Viana de Santana**

Título: "A DATA SCIENCE APPROACH TO LATTES CV DATA ANALYSIS"

Aprovado (a) pela Banca Examinadora
em cumprimento ao requisito exigido para
obtenção do Título de **Mestre** em
Computação Aplicada

Dr. Nandamudi Lankalapalli Vijaykumar

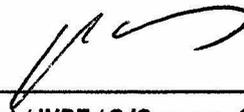


Presidente / INPE / SJC Campos - SP

() Participação por Video - Conferência

Aprovado () Reprovado

Dr. Rafael Duarte Coelho dos Santos



Orientador(a) / INPE / SJC Campos - SP

() Participação por Video - Conferência

Aprovado () Reprovado

Dr. Gilberto Ribeiro de Queiroz

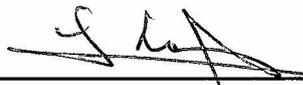


Membro da Banca / INPE / São José dos Campos - SP

() Participação por Video - Conferência

Aprovado () Reprovado

Dr. Ezzat Selim Chalhoub



Convidado(a) / . / SJC Campos - SP

() Participação por Video - Conferência

Aprovado () Reprovado

Este trabalho foi aprovado por:

() maioria simples

unanimidade

São José dos Campos, 20 de setembro de 2018

“They did not know it was impossible so they did it”.

MARK TWAIN

To my family.

ACKNOWLEDGEMENTS

To my advisor Rafael Duarte Coelho dos Santos, for the precious time dedicated to the orientation of this work.

To my wife, Elaine, for the encouragement and for the constant company on this journey in the easy days and the difficult days.

To my daughter Cora, who has not even completed one year of life, and already gives me so many joys. One day she will read this paragraph and, who knows, be happy too.

To my parents and family for the cheer, encouragement and presence.

To the CAP colleagues, walking parallel paths towards academic excellence.

To INPE, for supporting the academic training of so many researchers.

Some nations have monuments to pay homage to "The Unknown Soldier" – individuals who played their important part in that country's history. Likewise, there are other people that were part of the story of this work and that haven't been thanked nominally. To them, inspired by these countries' homage, I leave a thank you to "The Unknown Dear Ones".

ABSTRACT

The Lattes Platform is the *de facto* database of the Brazilian academic community. This web-based platform was created by the Brazilian National Council of Technological and Scientific Development (CNPq) and is updated by the researchers and students themselves, being of great value not only to store information about publications and other academic data about the users and their evaluation by the community but also for analysis of its data in different applications, such as to create reports, to evaluate research groups, higher-level educational programs and institutions. All data on the Lattes CV is public to a certain extent: CVs can be downloaded provided one knows the full name of the person of interest or its Lattes platform unique identifier.

Also, there are no native tools on the platform that allow specific analysis of groups of researchers and students; one must either browse or download a set of CVs and extract the required data from that set for posterior analysis.

This work intends to fill this gap by presenting a tool that processes and cleans up a Lattes CVs data set, that was developed with focus on users with little knowledge of programming. In this work we present the reports that this tool generates and that are related to Exploratory Data Analysis – such as reports generated with Lattes CV data – word clouds and graphs that exhibit relationship between researchers. This work also discusses extensions of this tool capabilities with unsupervised classification algorithms, showing its integration with artificial intelligence methods.

Keywords: Data Analysis. Data Science. Lattes Platform. Bibliometrics. Artificial Intelligence.

UMA ABORDAGEM DE CIÊNCIA DE DADOS PARA ANÁLISE DE DADOS DE CURRICULUM LATTES

RESUMO

A Plataforma Lattes é uma das principais bases de dados da comunidade acadêmica brasileira. Esta plataforma web foi criada pelo Conselho Nacional de Desenvolvimento Tecnológico e Científico (CNPq) e é atualizada pelos próprios pesquisadores e alunos, sendo de grande valor não só para a documentação das publicações e outros dados acadêmicos sobre os usuários e sua avaliação pela comunidade, mas também para a análise de seus dados em diferentes aplicações, por exemplo, para criar relatórios, avaliar grupos de pesquisa, programas educacionais de nível superior, instituições etc. Para realizar estas análises, os currículos devem ser baixados a priori. Todos os dados sobre o CV Lattes são públicos até certo ponto: para baixar os currículos é necessário conhecer o nome completo da pessoa de interesse ou o identificador exclusivo da plataforma Lattes.

Além disso, não há ferramentas nativas na plataforma que permitam a análise específica de grupos de pesquisadores e estudantes. Assim, deve-se fazer o download de um conjunto de Curriculum Lattes e extrair os dados requeridos desse conjunto.

Este trabalho pretende preencher essa lacuna através de uma ferramenta que processa e limpa o conjunto de dados Lattes CVs, permitindo seu uso por usuários com pouco conhecimento de linguagens de programação. São apresentados os relatórios que esta ferramenta gera e que estão relacionados à Análise Exploratória de Dados - como relatórios gerados com dados de Lattes CV - nuvens de palavras e gráficos que exibem relação entre pesquisadores. Também é discutida a extensão dessa ferramenta com algoritmos de classificação não supervisionados, mostrando sua integração com métodos de inteligência artificial.

Palavras-chave: Análise de Dados. Ciência de Dados. Plataforma Lattes. Bibliometria. Inteligência Artificial.

LIST OF FIGURES

	<u>Page</u>
1.1 Investments in R&D in GDP % per country, with Private Sector participation.	1
1.2 Total Investments in Scholarships and Promotion of Research.	2
1.3 Unemployment rates and average salary as a function of educational attainment.	3
2.1 Example of an XML document.	12
2.2 Example of a XSD document.	13
2.3 The Lattes CV XSD document.	14
2.4 Example of a XPath code in Python.	16
3.1 The Data Science Process.	22
3.2 Reproducibility spectrum of a scientific publication.	23
3.3 Example of a Jupyter Notebook window running the LattesLab tool.	25
4.1 Total Investments in Scientific Initiation Scholarships in BRL millions.	30
4.2 Number of Years of Scientific Initiations per Researcher.	32
4.3 Number of Months Since Last Lattes CV Update.	34
4.4 Researchers Nationality	35
4.5 Academic Level of the Researchers	36
4.6 Distribution of Graduation of Researchers per year.	37
4.7 Distribution of Master's Degrees of Researchers per year.	37
4.8 Distribution of Doctorate Degrees of Researchers per year.	38
4.9 Distribution of Post-Doctorate Degrees of Researchers per year.	38
4.10 Publication history of a researcher, obtained through Lattes CV Data.	39
4.11 Publication history of the group of researchers.	40
4.12 Word Cloud of the words in the Lattes CV summaries of all researchers.	41
4.13 Word Cloud of the words in titles of publications of the Lattes CV of a given researcher.	42
4.14 Word Cloud of the words in titles of publications of the Lattes CV of all researchers.	43
4.15 Collaboration Network of the Researchers.	45
5.1 Centroid coordinates: twenty-cluster model.	52
5.2 Centroid coordinates of the second experiment: twenty-cluster model.	54

5.3	Centroid coordinates of the second experiment and its associated data points: twenty-clusters model.	55
5.4	Centroid coordinates of the second experiment: three-clusters model. . .	56
5.5	Centroid coordinates of the second experiment and its associated data points: three-clusters model.	57
5.6	Number of data points associated with each neuron of the 20 x 20 SOM model.	59
5.7	Distribution of data points through each neuron of the 20 x 20 SOM model.	60
5.8	Number of data points associated with each neuron of the 3 x 3 SOM model.	61
5.9	Distribution of data points through each neuron of the 3 x 3 SOM model.	62
5.10	Evolution of SOM error for 3 x 3 neuron models and 5000 training epochs.	63
5.11	Evolution of error for 3 x 3 neuron models and 100000 training epochs. .	63
6.1	Mathematics Genealogy Tree.	68
6.2	Fuzzy C-Means model with 20 centroids and publications spanning through five years.	69

LIST OF TABLES

	<u>Page</u>
1.1 INPE's Postgraduate Courses Evaluation.	5
5.1 Error as a function of the number of clusters.	51
5.2 Error as a function of the number of clusters for the second experiment.	53
5.3 Error as a function of SOM neuron matrix dimension.	58

LIST OF ABBREVIATIONS

BMU	–	Best Matching Unit
BRL	–	Brazilian Reais (currency)
CAPTCHA	–	Completely Automated Public Turing test to tell Computers and Humans Apart
DSL	–	Domain-Specific Language
DOI	–	Digital Object Identifier
EDA	–	Exploratory Data Analysis
GDP	–	Gross Domestic Product
GUI	–	Graphical User Interface
HTML	–	Hypertext Markup Language
IDE	–	Integrated Development Environment
INPE	–	Instituto Nacional de Pesquisas Espaciais
MODS	–	Metadata Object Description Schema
OLAP	–	Online Analytical Processing
SID	–	Serviço de Informação e Documentação
SOM	–	Self-Organizing Map
SPG	–	Serviço de Pós-Graduação
TDI	–	Teses e Dissertações Internas
XML	–	eXtensible Markup Language
XSD	–	XML Schema Definition
XSLT	–	eXtensible Stylesheet Language Transformations

CONTENTS

	<u>Page</u>
1 INTRODUCTION	1
1.1 Research and Technology	1
1.2 INPE's Role	3
1.3 Research Productivity	5
1.4 Objectives	6
2 THE LATTES CURRICULUM VITAE	9
2.1 Data Extraction	10
2.2 Data Reliability	11
2.3 Structure of a Lattes CV	12
2.4 Lattes CV XSD	14
2.5 Lattes CV Data Extraction	15
2.5.1 The XPath Language	15
3 The LattesLab Library	17
3.1 Related Work	17
3.2 Data Science Approach	20
3.3 The LattesLab	23
3.4 Deployment	26
3.4.1 Python IDE Deployment	26
3.4.2 Jupyter Notebook Deployment	27
4 THE LATTES CV COLLECTION	29
4.1 INPE's Scientific Initiation Program	29
4.2 Lattes CV Collection	30
4.3 Exploratory Data Analysis	31
4.3.1 Scientific Initiation Scholarships per Student	32
4.3.2 Lattes CV Last Update in months	33
4.3.3 Other Lattes CV Data	35
4.3.4 Wordclouds	40
4.3.5 Graphs	44
5 QUANTITATIVE LATTES CV ANALYSIS	47

5.1	Introduction	47
5.1.1	Clustering Algorithms	47
5.1.2	Lattes CV Publication Data	47
5.2	Clustering Algorithms	49
5.2.1	Fuzzy C-Means Algorithm	49
5.2.2	Self-Organizing Map (SOM) Algorithm	50
5.3	Quantitative Results	51
5.3.1	Fuzzy C-Means Results	51
5.3.2	Self-Organizing Map (SOM) Results	57
5.4	Discussion	61
6	CONCLUSION AND FUTURE WORKS	65
6.1	Contributions	65
6.2	Conclusion	66
6.3	Publications	67
6.4	Future Work	67
	REFERENCES	71

1 INTRODUCTION

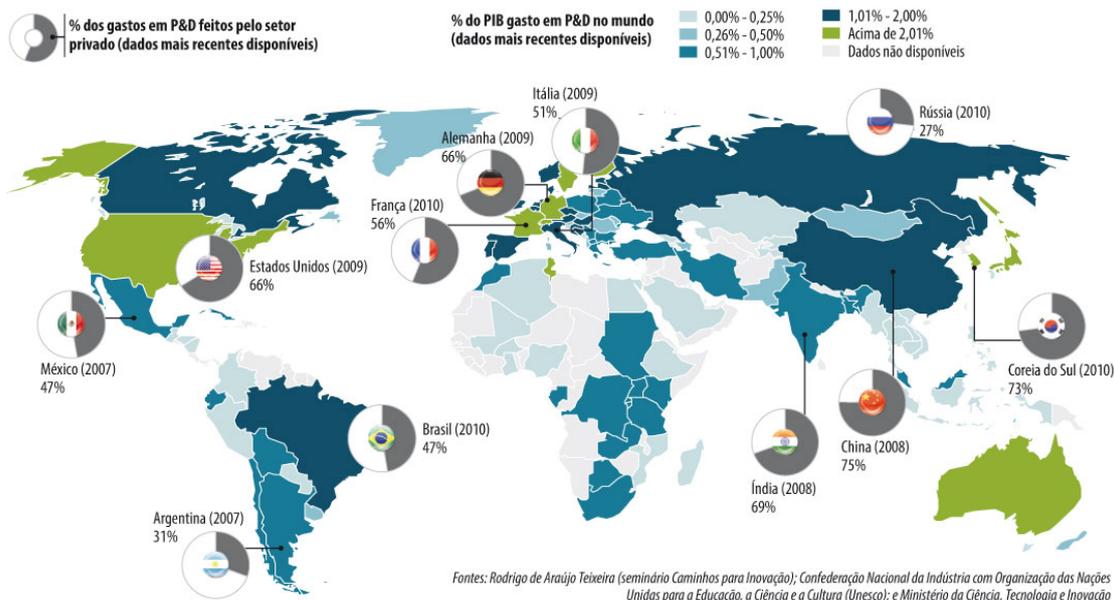
1.1 Research and Technology

The volume of investments governments make in research and development (R&D) is a strong indicative of how important the education of students and the generation of knowledge is relevant for that country. Figure 1.1 shows, for some countries, the fraction of the GDP that is invested in Research and Development and also which percentage of these investments are made by their government and which part is made by the private sector.

Figure 1.1 - Investments in R&D in GDP % per country, with Private Sector participation.

Empresas arcam com até 75% dos investimentos em P&D no mundo. No Brasil, Estado paga a metade

América do Norte, Ásia e Europa concentram cerca de 90% dos gastos em pesquisa e desenvolvimento. Nesses continentes, o setor privado responde pela maior parte dos projetos inovadores, ainda que subsidiados ou subvencionados pelos governos



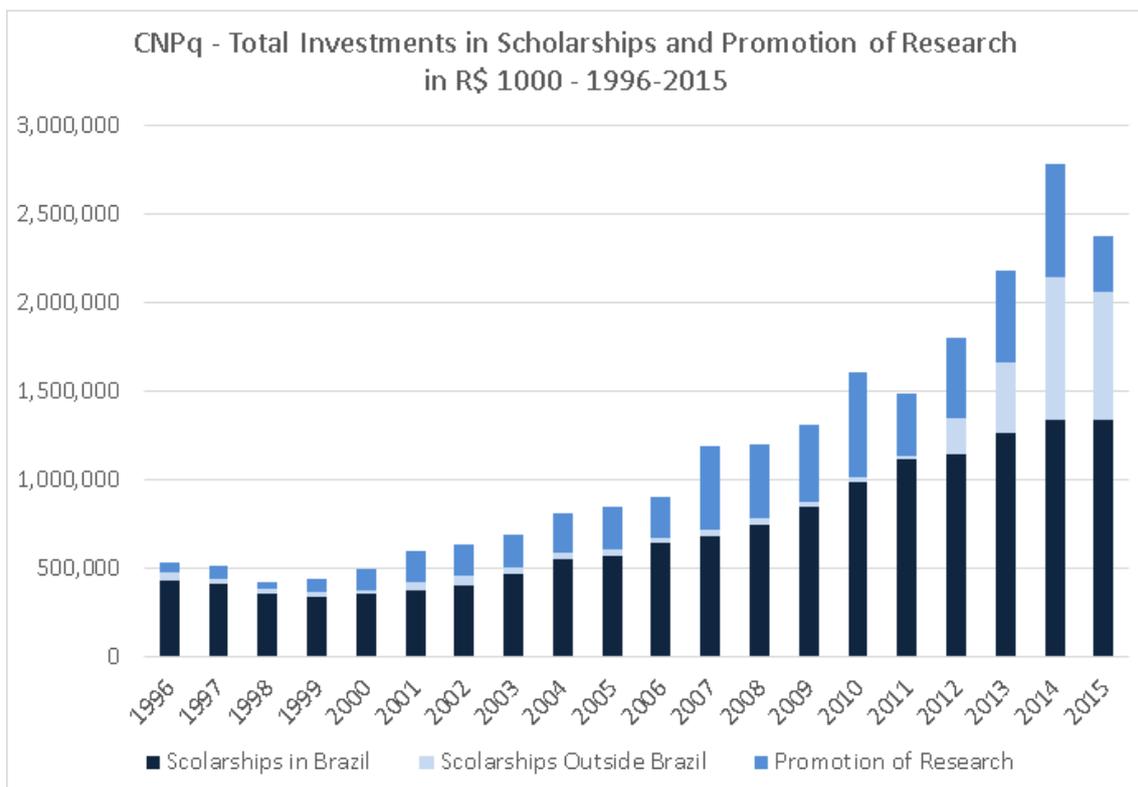
SOURCE: Brasil. Senado Federal (2012).

These investments in R&D have the potential to generate results for research universities: either intangible results – like academic training of their alumni and generation of intellectual property such as patents – and tangible – such as the financial results that come through these patents. These research universities may also associate with

companies and aim for the common goal of achieving innovation and scientific discoveries. Discoveries which these companies may fund to increase their own profits (UNIVERSITY WORLD NEWS, 2018).

In Brazil, these investments in R&D have grown steadily up to the year 2015. From 1995 to 2015, the investments in scholarships and for research promotion have risen from almost 500 million reais to 2,300 million reais - growing over four times (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ, 2018d). These values are found in Figure 1.2.

Figure 1.2 - Total Investments in Scholarships and Promotion of Research.



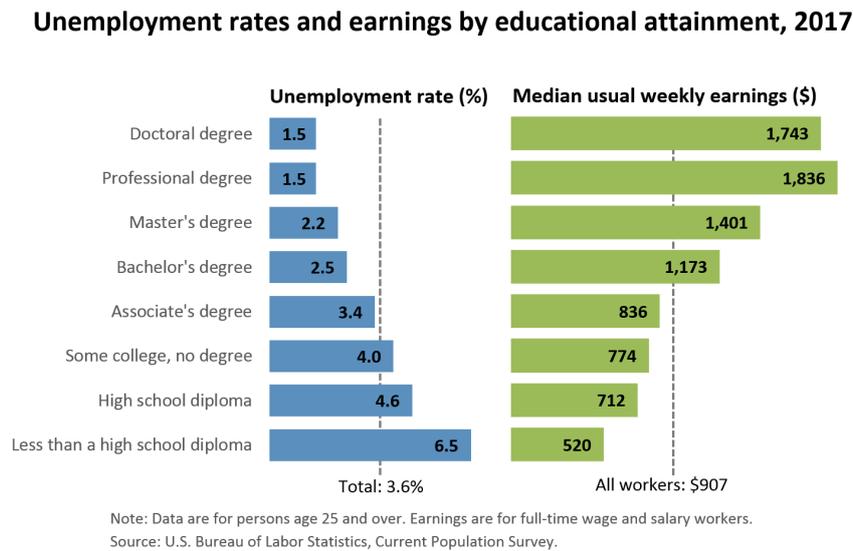
SOURCE: Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (2018e).

After 2015, however, the investments in education have diminished (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ, 2018c). This decrease is correlated to a fall in Brazil's GDP in 2015 and 2016 (WORLD BANK,

2018).

The academic environment is not only profitable for academics invested in applying their knowledge to the expansion of scientific knowledge frontiers. Statistical data from the United States' Bureau of Labor Statistics (BUREAU OF LABOR STATISTICS, 2018) shows that, the higher the education level of an individual, the lower the unemployment rate of that individual and the higher their weekly earnings - with the exception of owners of a Doctoral Degree, which earn less in average than those with a Professional Degree. Figure 1.3 shows this relationship.

Figure 1.3 - Unemployment rates and average salary as a function of educational attainment.



SOURCE: Bureau of Labor Statistics (2018).

1.2 INPE's Role

The Brazilian government has assumed an important role of promoting investments in Research and Development. According to the data in Figure 1.1, government was responsible for more than the half of all investments in R&D in Brazil, while in other countries up to 75% of the R&D budget comes from the private sector

(BRASIL. SENADO FEDERAL, 2012).

In this scenario, the Brazilian National Institute for Space Research (*Instituto Nacional de Pesquisas Espaciais – INPE*) is a significant player. The institute is a research unit of the Ministry of Science, Technology, Innovation and Communication responsible for developing scientific research and technological knowledge. Its objectives are:

- To expand and consolidate science, technology and innovation competencies in the space and terrestrial environment.
- To develop scientific and technological leadership in the space and terrestrial environment.
- To expand and consolidate scientific knowledge in weather and climate forecasting and global environmental change.
- To consolidate itself as a singular institution in the development of satellites and space technologies.
- Promote a space policy for industry, with the objective of meeting the developmental needs of space services, technologies and systems.
- Strengthen its own institutional relationships at national and international levels.
- Provide adequate infrastructure for scientific and technological development.
- Identify and implement a managerial and institutional model, adequate to the institute's challenges.

INPE achieves its scientific and personnel training goals through its postgraduate courses. It has currently seven programs with Masters and/or Doctorate degrees, all considered of good or excellent performance by CAPES (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - UFRN, 2018). These concepts are detailed in Table 1.1.

Table 1.1 - INPE's Postgraduate Courses Evaluation.

PROGRAM NAME	PROGRAM CODE	Masters Degree Rate	Doctorate Degree Rate
ASTROFÍSICA (ASTROPHYSICS)	33010013010P4	4	4
CIÊNCIA DO SISTEMA TERRESTRE (SCIENCE OF THE TERRESTRIAL SYSTEM)	33010013011P0	-	6
COMPUTAÇÃO APLICADA (APPLIED COMPUTING)	33010013002P1	5	5
ENGENHARIA E TECNOLOGIA ESPACIAIS (SPACE ENGINEERING AND TECHNOLOGY)	33010013009P6	4	4
GEOFÍSICA ESPACIAL (SPACE GEOPHYSICS)	33010013008P0	6	6
METEOROLOGIA (METEOROLOGY)	33010013003P8	6	6
SENSORIAMENTO REMOTO (REMOTE SENSING)	33010013005P0	7	7

The evaluation generates scores from 1 through 7. Courses evaluated as 1 and 2 are not recommended by CAPES. Courses with a score of 6 and 7 have a performance equivalent to that of the most important international teaching and research centers. A score of 5 is given for courses with a high level of academic performance; a score of 4 is considered a good performance and a score equal to 3 is given for courses with the minimum quality standards.

1.3 Research Productivity

From the discussion in section 1.1, it is possible to conclude that investments in research and higher education can result not only in knowledge generation, but can also generate a higher employment rate and salaries for the individuals pursuing that higher education.

These investments are found directly or indirectly in research institutes and universities promoting research and higher education programs. Therefore, one way to verify the effectiveness of these investments is to assess the results of these institutes and their faculty members.

In this assessment, it is not possible to say that the productivity of a researcher, or group of researchers, is correlated to its teaching effectiveness (CENTRA, 1981). However, the scientific output of a researcher is often related to a productivity score, that may be used to evaluate how productive a researcher is.

One productivity score used to evaluate scientific production is the h-index (HIRSCH, 2005). If the h-index of a given researcher is n , it means that researcher has n articles published throughout his or hers career that have at least n citations each one. A historical ranking of the highest h-indexes has Sigmund Freud on the top of the list with an index of 272 (RANKING WEB OF UNIVERSITIES, 2018).

Evaluation of researchers and students productivity, either individually or in groups, is an important task for universities, research centers, and funding agencies, and also for the students and researchers themselves: students may be interested in knowing more about the achievements, research areas and experiences of prospective advisers. Teachers and admission deans may also want to know about the academic history of candidates to a graduate program, for example. The usual method to evaluate academics and students is by analysis of the achievements and publications listed on his or hers curriculum vitae (CV). However, as much as there is access to academic data from researchers and researcher centers (UNITED NATIONS - UN, 2018) (FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO – FAPESP, 2018) (GOOGLE, 2018), the challenge of transforming this data into reports, graphics and data products which can be used to guide decisions – of students, researchers and administrators – remains.

1.4 Objectives

The objectives of this dissertation are:

- To propose a tool to evaluate science production in the Brazilian scientific environment. For that, the Lattes Platform and its Lattes CV files are used, since they are the standard way researchers and their work are presented in Brazil. For the purpose of this work, Lattes CVs of researchers related to INPE were used, and the relationship of these researchers with INPE is clarified.
- To justify the need of a new tool for this analysis. For that, an analysis of tools that were developed and used previously to retrieve information from the Lattes Platform – and that used this data to generate their own

results – is performed.

- To present the capabilities of the proposed tool. Since data related to scientific production is the object of this work, a Data Science approach to gather, filter and present this data is proposed.
- With a Data Science approach, use of artificial intelligence algorithms to extract results from Lattes CV data is viable. Therefore, a data classification problem is proposed and some of its results are discussed in this work.
- Discuss the possibility of these analyses being expanded for uses different than those applications presented in this work.

To accomplish these objectives, this dissertation is organized in the following way:

- Chapter 2 discusses the database used to retrieve the researchers' information - the Lattes CV platform. Its potentials and limitations are presented and discussed throughout the chapter.
- Chapter 3 describes the tools that were developed in the past to retrieve data from the Lattes Platform, and that used this data to analyze the researchers' information. With the presentation of these tools, a new tool is proposed, and what functionalities this tool must present to be a valuable asset, considering the current state of the Lattes Platform.
- Chapter 4 explains the data contained in the Lattes CVs, and some problems associated with this data. It contains also qualitative analysis of the Lattes CV set that is used in this work and Exploratory Data Analysis of its data.
- Chapter 5 brings an extension of the analysis of the Lattes CV data by applying unsupervised learning algorithms to analyze the publication history of the researchers, proposing possible answers to the question: are there mean profiles of publications to be expected from researchers?
- Finally, Chapter 6 concludes the dissertation and its achievements and proposes future works derived from its results.

2 THE LATTES CURRICULUM VITAE

As stated in section 1.3, a researcher's CV is a typical source of information about them. There, it may be possible to find their history of publications, peer network, association with universities and research centers and other relevant information. Therefore, to attain the objectives of this dissertation, it is necessary to access the CVs of the group of researchers whose data is to be analyzed. The Lattes Platform is where these CV are found.

The Lattes Platform¹ (named in honor of César Lattes, a Brazilian physicist) is an online system maintained by the Brazilian National Council of Technological and Scientific Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*, CNPq). It provides a unified interface to a database used to collect, store and process academic achievements information. Any researcher can create and maintain his or hers own Lattes CV, using the taxonomy, formats and fields defined by the platform. The standardization of some of its fields and categories makes it easy to fill the forms that feed the database.

The CV Lattes is the Brazilian standard adopted for both researchers and institutions to publish scientific related data. It is available at the CNPq website and its access is open to the public. The network is maintained by the Brazilian government.

The Lattes Platform is also used by CNPq to generate reports about the current status of the academic production of the researchers and students, and to evaluate applications to several different types of grants. The data on the platform is also used by other government funding agencies and by the Ministry of Education, for evaluation of the production of professors and students in graduate programs.

The Lattes Platform public interface is a web-based system that allows the edition of the CVs by its owner and the search and retrieval of the CVs by anyone that knows either the researchers' names or IDs (a 16-digit unique identifier). In the end of 2016 there were more than 3.500.000 CVs stored in the database². Of those, almost 1.500.000 were CVs of students.

Data in the Lattes Platform can also be used for other academic purposes: analysis of academic indicators' evolution (PEREZ-CERVANTES *et al.*, 2012), identification of communities based on similar interests or collaborations (MENA-CHALCO *et al.*, 2014;

¹<http://lattes.cnpq.br>

²<http://estatico.cnpq.br/painelLattes/>

ARAÚJO et al., 2014; ALVES et al., 2011b), changes of research areas based on the publications records, etc. Analysis considering groups of researchers or students can be done in different scales, from the whole academic community to small groups, such as researchers in a group or students in a college. But although the Lattes CV data is considered public (being provided by the researchers and students themselves), retrieval of the data is limited: it is possible to download an individual CV as an XML file with all the data entered on that CV, but it is not possible to retrieve more than one CV at a time. This makes it hard to perform some specific types of analysis that requires the extraction of certain categories from several CVs at once.

2.1 Data Extraction

Acknowledging the interest in the data it encompasses, the platform allows the possibility to download each CV Lattes individually. However, to perform this download, one must know the Lattes identification of the researcher of the desired Lattes CV and there is also a CAPTCHA test. These two characteristics prevent an automation of the download process, and it can only be conducted manually.

Previous research that used automated tools (MENA-CHALCO; CESAR JUNIOR, 2009) stopped working after migration of the Lattes Platform to an XML based environment, and implementation of the mentioned CAPTCHA test. The loss of automation capability has impaired the use of the Lattes database as a ready-to-use database, and all work related to this data has now to be preceded by the download of each CV Lattes, one by one.

CNPq recognizes the usefulness of the Lattes platform data for each institution and allows the access of each institution to the Lattes CVs of its associated researchers (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ, 2017). There are two possibilities for this access:

- **Mirroring:** Public research centers may request this access modality. It consists of the integral availability of the data of the Platform Lattes so that it can mirror the data properly inside the institution. In this modality, the mirrored Lattes data is updated daily. Any institution aspiring to mirror the Lattes database must send a letter – signed by its highest leader – explaining the reasons for the request to the CNPq President.
- **Extraction of CV data and Research Groups:** it is available to all research and innovation institutions in Brazil. An interested institution

must send a letter to the CNPq President, signed by its Chief Executive Officer, containing the justification of the request and destination of the extracted data. Institutions that have a Protocol of Technical Cooperation signed with CNPq already have the right of access to the data.

In any of the two cases presented, the data obtained is always supplied in the standard XML format defined by the Community Conscientias, responsible for standardization of the Lattes Platform information.

2.2 Data Reliability

There is also the question of the reliability of information published on the Lattes Platform. Because it is an open platform, despite bringing transparency and accessibility to the dissemination of researchers and their work, there is the opportunity to publish unlikely information in this easy-to-access platform. To mitigate this issue, the Lattes Platform may include the Digital Object Identifier (DOI) code for the listed publications and the cross-reference to the CAPES thesis database (TORRESI *et al.*, 2009). The insertion of DOI, of course, does not eliminate errors such as an inadequate citation of authors and co-authors – often making disambiguation impossible.

Besides, researchers keep on filling in their Lattes CV even when they have doubts about the correct way to fill in the CV fields (MARQUES *et al.*, 2007). This behavior, and the absence of a manual or short course on how to fill in the Lattes CV correctly, leads to creation and propagation of incorrect data which, when consulted, may lead to incorrect conclusions.

Regardless of the obstacles associated with its use, Lattes Platform remains a free, rich and interesting source of information to the scientific community.

In possession of a group of Lattes CV – regardless of the way it was acquired – it is possible to perform data analysis addressing several questions, such as:

- How many papers a given department is producing per year?
- What is the academic degree of a group of students of one university department?
- How many scientific research scholarships were granted by the department in each year?

2.3 Structure of a Lattes CV

The questions raised previously can all be answered through data found in Lattes CVs, and to the fact that this data can be procedurally accessed thanks to the Lattes CV files organization.

Currently, each Lattes CV is presented in XML format. According to the W3 foundation (W3SCHOOLS, 2018a) XML, which stands for eXtensible Markup Language, is a markup language – much like HTML – designed to store and transport data and to be self-descriptive. The XML language contain tags that form the structure of the XML document and metadata associated with each tag. Tags and metadata can be optional or required, which is defined by the XML Schema Definition (XSD).

An example of an XML is presented in the Figure 2.1. An example of XSD associated with the previous XML document is presented in Figure 2.2.

Figure 2.1 - Example of an XML document.

```
<?xml version="1.0" encoding="UTF-8"?>
- <badge status="active">
  <name>Howard Kierkegaard</name>
  <IDnumber>669822</IDnumber>
  <birthday>06/09/1981</birthday>
</badge>
```

SOURCE: Example of XML written by the author.

Figure 2.2 - Example of a XSD document.

```
<?xml version="1.0"?>
- <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  - <xs:element name="badge">
    - <xs:complexType>
      <xs:attribute name="status" default="active" type="xs:string"/>
      - <xs:sequence>
        <xs:element name="name" type="xs:string"/>
        <xs:element name="IDnumber" type="xs:string"/>
        <xs:element name="birthday" type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

SOURCE: Example of XSD written by the author.

In the XSD example in Figure 2.2, it is possible to see two types of data found in an XML document:

- Elements: an element comprises every piece of data found between its start and end tags, including the tags themselves. An element may be composed of attributes, data, and/or other elements. In the example above, the element “badge” is formed by an attribute “status” and three other elements: “name”, “IDnumber” and “birthday”. The element “name” contains only data, in the form of text.
- Attributes: attributes also contain data, but are found within the starting tag of an element. Different than elements, attributes cannot contain multiple values, and do not accept “child” entities – like elements do.

There are no rules for use of an Element or Attribute in XML, although meta-data (i.e., data about the data) should be stored in attributes, whereas the XML document data should be stored in elements (W3SCHOOLS, 2018b).

There are also other entities that may be found in an XML file, such as comments, document nodes, namespace processing-instruction and text. These entities, however, are not relevant for the Lattes CV analysis, since they are not found in a

Lattes CV XML document. In these documents, the information considered relevant for this analysis is found in elements and attributes.

2.4 Lattes CV XSD

The Lattes CV XSD (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ, 2018b) lists all elements and attributes that can be found in a Lattes CV XML document. Due to the complex data contained on the Lattes CV, it is no surprise that the XSD houses a big list of possible elements and attributes. Some of the first elements XML elements are listed on Figure 2.3.

Figure 2.3 - The Lattes CV XSD document.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
  xmlns:xs="http://www.w3.org/2001/XMLSchema">
  - <xs:element name="CURRICULO-VITAE">
    - <xs:complexType>
      - <xs:sequence>
        <xs:element ref="DADOS-GERAIS"/>
        <xs:element ref="PRODUCAO-BIBLIOGRAFICA" minOccurs="0"/>
        <xs:element ref="PRODUCAO-TECNICA" minOccurs="0"/>
        <xs:element ref="OUTRA-PRODUCAO" minOccurs="0"/>
        <xs:element ref="DADOS-COMPLEMENTARES" minOccurs="0"/>
      </xs:sequence>
      <xs:attribute name="SISTEMA-ORIGEM-XML" use="required"/>
      <xs:attribute name="NUMERO-IDENTIFICADOR"/>
    - <xs:attribute name="FORMATO-DATA-ATUALIZACAO" default="DDMMAAAA">
      - <xs:simpleType>
        - <xs:restriction base="xs:NMTOKEN">
          <xs:enumeration value="DDMMAAAA"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="DATA-ATUALIZACAO"/>
  - <xs:attribute name="FORMATO-HORA-ATUALIZACAO" default="HHMMSS">
    - <xs:simpleType>
      - <xs:restriction base="xs:NMTOKEN">
        <xs:enumeration value="HHMMSS"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:attribute>
</xs:element>
</xs:schema>
```

SOURCE: Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (2018b)

From Figure 2.3, it is possible to understand part of the structure of a Lattes CV XML. For example, it is expected that the XML file contains an element “CURRICULO-VITAE”, which itself contains the elements “DADOS-

GERAIS”, “PRODUCAO-BIBLIOGRAFICA”, “PRODUCAO-TECNICA”, “OUTRA-PRODUCAO” and “DADOS-COMPLEMENTARES”. The same element “CURRICULO-VITAE” has also the attributes “SISTEMA-ORIGEM-XML”, “NUMERO-IDENTIFICADOR” and “FORMATO-DATA-ATUALIZACAO”.

Through the knowledge acquired from the Lattes CV XSD, the Lattes CV data arrangement is unveiled. With this knowledge at hand, it is possible to create automated routines to extract the data from the Lattes CV files. The tool that is used to extract data from XML files – such as the Lattes CVs – is described in section 2.5.

2.5 Lattes CV Data Extraction

On the previous sections, the Lattes CV is characterized as an XML document, and the structure of this document is described through its XSD. In possession of this information, the information of the Lattes CV can be extracted through the XPath language.

2.5.1 The XPath Language

The XPath language is used to manipulate sections of an XML document, its elements and attributes. It uses a path-like syntax to navigate through the XML document, retrieving, assessing and editing its information.

For this language, the XML document is treated as a tree of nodes. These nodes can be the XML elements, attributes, and other entities.

XPath is implemented in a series of programming languages. In Python language, the library `ElementTree` has limited support to XPath language commands. An example of a XPath routine implemented in Python language is presented in figure 2.4.

It is possible to see in figure 2.4 that:

- Line 8 of the code imports the `ElementTree` library;
- Line 19 of the code contains a command to parse the file obtained from the previous commands and stores the parsed XML file in the variable *tree*;
- The code in line 20 imports the parsed XML root into the variable *root*. It is possible to navigate this *root* variable and look for elements, attributes,

Figure 2.4 - Example of a XPath code in Python.

```
1 def lattes_owner(cfolder, filename):
2     """ Returns the name of the owner of the Lattes CV found in arg filename
3     Args:
4         filename: complete path (folder + file) of the file where the Lattes
5         CV is found. The file must be a zip file containing a XML file.
6     """
7     import zipfile
8     import xml.etree.ElementTree as ET
9     import os
10
11     folder = os.path.normpath(cfolder)
12     rightfile = os.path.join(folder, filename)
13
14     #opens the Lattes CV file
15     archive = zipfile.ZipFile(rightfile, 'r')
16     cvfile = archive.open('curriculo.xml', 'r')
17
18     #initializes xpath
19     tree = ET.parse(cvfile)
20     root = tree.getroot()
21
22     #get the cv owner name
23     cvowner = root[0].attrib['NOME-COMPLETO']
24     return cvowner
```

SOURCE: Screen of a notepad app containing Python XPath code.

in different "levels" of the tree - always according to the Lattes XSD;

- The code is concluded when the function retrieves the name of the Lattes CV file owner from the attribute "NOME-COMPLETO", found in the element *root[0]*.

With the knowledge that the Lattes CVs' data can be mapped according to its XSD, and that the XPath language makes possible to retrieve this data from the Lattes CV's XMLs, the tools that are used to perform the analysis in this publication can be analyzed in Chapter 3.

3 The LattesLab Library

In this section, the LattesLab Python library – which is used to retrieve the data from the Lattes CVs and perform the data analysis – is introduced. Also, a review of the bibliography related to Lattes CV associated tools is presented in section 3.1.

3.1 Related Work

Access to Lattes CVs data is a requirement for several different types of analysis, and often these analyses must be done considering collections and not individual CVs. Considering this need, several different tools were created in the past to process and analyze data from the Lattes Platform.

One of the first tools that allowed extraction of information from a collection of Lattes CVs is *scriptLattes* (MENA-CHALCO; CESAR JUNIOR, 2009). This tool extracted data from the Lattes CVs of groups of researchers, creating reports, maps, graphs and other information from the collections of CVs. The tool could be deployed as a local application on computers running Linux, and its authors made the tool open-source, so other researchers could use it. Other research groups used *scriptLattes* as basis to create different types of analyses (ALVES et al., 2016), (PEREZ-CERVANTES et al., 2012).

Initially *scriptLattes* was designed to download the Lattes CVs data from CNPq’s servers as HTML files. However, with later adoption of a CAPTCHA (“Completely Automated Public Turing test to tell Computers and Humans Apart”) access control system, automatic download was hindered, so the tool was modified to use a local set of files that must be downloaded in advance.

scriptLattes is probably the most referenced Lattes-CV-related tool in the bibliography. It contains several features, but relies on parsing the data from the HTML version of the Lattes CVs, which has changed in the past and may change in the future. Reports and graphics are also pre-programmed, so extensions and different layouts must be programmed separately.

*LattesExtractor*¹ is a tool developed by CNPq that allows the download of several Lattes XML CVs in batches. Although it seems to solve part of the problem at hand, namely, how to obtain subsets of the data, it is not as open or as flexible: only registered organizations can retrieve data with this tool, and organizations can only

¹<http://lattesextractor.cnpq.br/lattesextractor/>

access data from their own organization. For example, a university may be able to download all the XML files with the CVs of its staff, teachers and students, but will not be able to download CVs of collaborators that don't currently work or study at the university. Similarly, when a student graduates and leaves the university, its CV is no longer available after graduation. This tool also does not perform any kind of analysis, providing only the XML files.

SUCUPIRA (ALVES et al., 2011b) was developed as a tool that allowed both the semiautomatic extraction of the XML files from the Lattes Platform and the creation of reports and graphics that could answer questions about collaboration between researchers, their geographical location, their scientific production and its evolution, etc. *SUCUPIRA* is a web-based application that uses a list of names of researchers or students, managed by the system's user, to download the Lattes CVs (as HTML files), parse them and create reports based on the data extracted from the CVs on that list.

The development of that tool was discontinued; also, changes on the Lattes Platform made it unusable in its current state: the structure of the HTML files changed, therefore parsers that could parse a version of the HTML generated by the platform had their usefulness restricted due to the new layout. Additionally, *SUCUPIRA* was written when access to the Lattes CVs was unhindered by CAPTCHAS – for some time the platform used simple CAPTCHAS that could be solved with tools such as Tesseract (KAY, 2007), but recently the CAPTCHAs were made more difficult to solve automatically.

SUCUPIRA used another tool developed by the same researchers: *LattesMiner* (ALVES et al., 2011a). *LattesMiner* is a Domain-Specific Language (DSL) implemented as a set of Java classes that allows the manipulation of data on a set of Lattes CVs, defined by a programmer, with modules for data discovery (association of names and IDs), data extraction (parsing of the HTML files corresponding to the CVs with regular expressions), storage of data in a local database, visualization and analysis tools. As with *SUCUPIRA*, *LattesMiner*'s development has stopped and changes on the Lattes platform rendered some aspects of the tool unusable.

Another tool that was designed to process data from the Lattes Platform is *XMLLattes* (FERNANDES et al., 2011), which converts the Lattes CVs from HTML to XML for further processing. This feature, however, is now unnecessary since the present version of the Lattes platform already exports the XML file of the curriculum vitae (although requiring CAPTCHAs for download of individual CVs).

As part of this research several papers related to analysis of Lattes CVs data (DIAMPIETRI et al., 2012), (ARAÚJO et al., 2014), (MENA-CHALCO et al., 2014), (PEREZ-CERVANTES et al., 2012) were reviewed. Most of these papers used a database with detailed information on academics, which was extracted from the Lattes Platform when it was possible to do so without the limitations imposed by the CAPTCHA currently in use. There were no references on whether that database was kept up-to-date.

There is also another set of Lattes CVs tools dedicated to ontology analysis. In Computer Science, an ontology is a representation of the relationship between elements that represent knowledge. There are several reasons to develop an ontology (NOY et al., 2001), most of them related to defining a knowledge, classify and analyze it and to enable its reuse. The elements in a Lattes CV may clearly apply to be subjects to ontology analysis, since there are relationships between researchers and their productions, the schools they have studied in, their advisers, and the other elements in a Lattes CV.

(BONIFACIO, 2002) proposes an ontology named *OntoLattes* which encapsulates the researchers Lattes CV data. This ontology was further extended by (NAKASHIMA, 2004), which converted it from a DAML-OIL format to an OWL format and used other types of classes and data, not only objects.

*SemanticLattes*² is another ontology-based tool that converts each Lattes CV XML into a Metadata Object Description Schema (MODS) file³. This file format is an XML-based bibliographic description schema developed by the United States Library of Congress' Network Development and Standards Office typically used in library applications.

*SOS Lattes*⁴ (GALEGO, 2013) is a more recent ontology tool that was developed over the *scriptLattes* and *SemanticLattes*. It uses the data processed by these tools to report inconsistencies between Lattes CVs, and generate reports from a group of Lattes CVs that is being analyzed by the tool. Since it is built on *scriptLattes*, it can also only be used in a Linux (or Windows virtual machines) environment.

²<https://github.com/arademaker/SLattes>

³<http://www.loc.gov/standards/mods/>

⁴<https://github.com/efgalego/SOSLattes>

3.2 Data Science Approach

Up to now, a number of Lattes CV exploring tools were listed. Some of them cannot currently be used as designed for the following reasons:

- The Lattes CV platform has updated its data publishing technology from HTML to XML. Therefore, all the tools that relied on that previous format distribution no longer work properly. It seems a trivial technical issue, but the tools that extracted information from the Lattes CVs formatted as HTML documents had to deal with complex HTML structures and had to detect, from the HTML content itself, the categories of information being extracted (e.g. articles in journals, conferences, titles, authors, etc.) while the data represented as XML is properly formatted and tagged with this information. Therefore, even though the XML platform restricted the use of the HTML-based tools, it improved the Lattes CV files structural elements, facilitating access to specific data inside these CV files and laying the foundation for development of new, more robust tools.
- Some of these tools were designed to automatically download a list of Lattes CVs from CNPq's site. This is not possible today due to the implementation of the CAPTCHA test, both to visualize and to download the Lattes CVs.
- Some of the solutions are only available on specific platforms - such as Linux - and require a specific setup before use.

Considering the status of the existing tools for Lattes CV analysis, a new tool can be introduced to extract, interpret, analyze and visualize the data in a simple but flexible way.

Any such tool must work with the current state of the Lattes Platform. For that, it must:

- Work with an offline set of Lattes CVs. Due to the current impossibility of automated batch download of Lattes CVs, the CVs must either be obtained manually or automatically through other authorized tools – such as *LattesExtractor*.
- Be able to transform the list of Lattes CVs' XML files into table-like data structures for further processing. To make this transformation, it is neces-

sary to know the structure of the Lattes CV XML, identify the parameters of interest and migrate these parameters to the data structures. This transformation is made easier since CNPq publishes the XML Schema Dictionary (XSD) of the Lattes CV files.

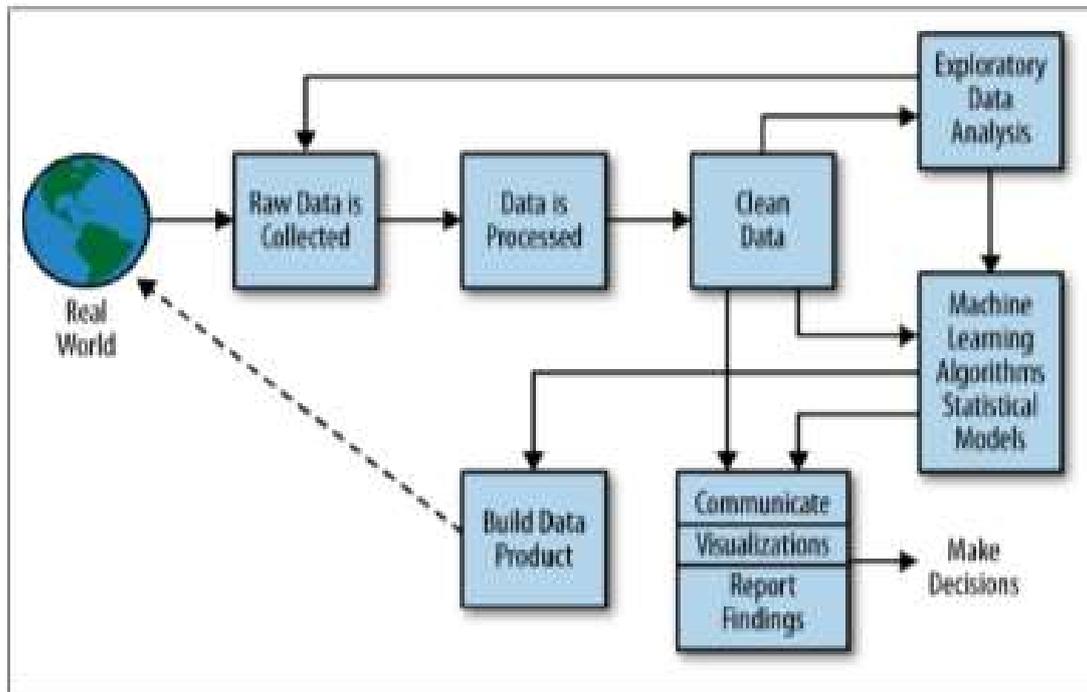
- Be agnostic with respect to the operating system used to run the tool. Each potential user has a limited number of resources and to require the user to learn a new programming language, install a new software or even a new whole operating system shall be avoided if the tool is expected to be widely used.
- Require no specialized knowledge for operation. In the same way that requiring a specific environment limits the utilization of the tool, if specialized knowledge is not required to use it, the tool could potentially be used by a higher number of individuals. At the same time the tool must be extensible so that advanced users could implement different features from the tool.
- The results produced by this tool should be reproducible by any user interested in analysis of a set of Lattes CVs. Reproducibility is ensured by the use of a common set of instructions that can be easily shared and build upon.

By designing a tool that follows these requirements it is possible to apply a Data Science process to the problem of analysis of collections of Lattes CVs. The data science process (SCHUTT; O'NEIL, 2013) is presented in Figure 3.1.

The requirements listed previously are directly related to the Data Science process: the first requirement makes reference to the collection of data, the second requirement is related to the processing and cleaning of data as well as storage of the “clean” data, so that it can be easily and transparently accessed. Even though the fourth and fifth requirements are not directly associated to the steps shown in the Data Science process (Figure 3.1), they serve as a guideline to ensure that the results attained by the tool are easy to acquire — and customize — and are also reproducible.

The fourth requirement listed on this section is also directly related to the concept of Exploratory Data Analysis (EDA), which intent is to allow the researcher to discover patterns on data by using visualization tools and statistics to understand the behaviors and characteristics of a given data set.

Figure 3.1 - The Data Science Process.



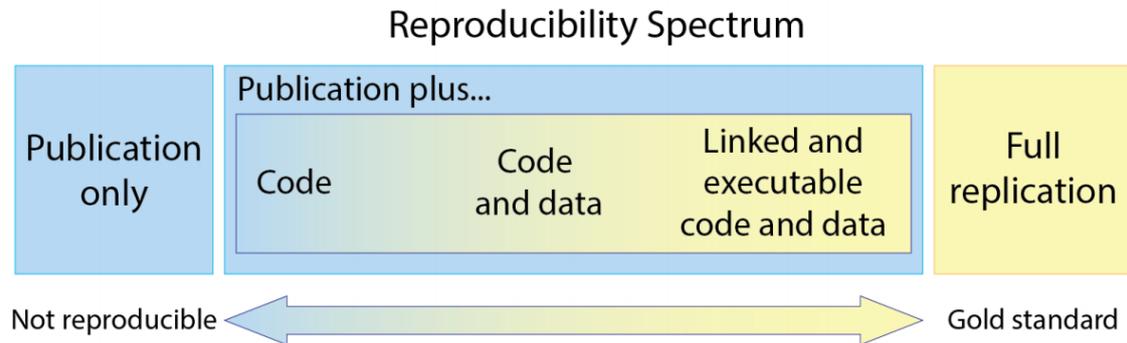
SOURCE: Schutt e O'Neil (2013).

Analysis of Lattes CVs can be done in different ways, using different metrics and algorithms. Considering that most of the analyses involves thematic groups of researchers (e.g. researchers in a specific area of knowledge, or professors and students of a specific department), it is very likely that methods and techniques applied to a particular analysis can be used in different contexts. A tool for the analysis of Lattes CVs collections must facilitate the reproduction of its results — since reproducible research is also a concept closely linked to Data Science.

According to (PENG, 2011), there is a spectrum of reproducibility of research, that goes from a non-reproducible result to a fully reproducible one (Figure 3.2).

Considering Figure 3.2, it is highly desirable that a tool performing Lattes CV Analysis allows the full reproduction of the data analysis experiments with the corresponding code being publishable and applicable to different sets of data of the same nature.

Figure 3.2 - Reproducibility spectrum of a scientific publication.



SOURCE: Peng (2011).

3.3 The LattesLab

In order to address the requirements listed on the section 3.2, it is proposed a software stack solution, based on concepts and principles of Data Science, to tackle the generic problem of analyzing Lattes CVs. This tool, named LattesLab, is based on the following components:

- A library that is able to scan a collection of Lattes CVs (stored as local files) and create a set of data frames (table-like structures) from that collection.
- A deployment mechanism for that library that allows its use, with minimal software installation requirements.
- A set of live documents that shows how to perform basic statistical analysis, visualizations and reports.

LattesLab is developed in the Python language's (VANROSSUM; DRAKE, 2010) widely used platform -- considering the size of its communities of developers and users. It is also one of the most used programming languages to solve Data Science problems due do the large amount of free and open analysis and visualization libraries. Other languages were considered for implementation -- a prototype was developed in Java, but the distribution of the library to use in other derived projects was considered complex. R (IHAKA; GENTLEMAN, 1996) was also considered: even though the R language is strongly supported by the community and is heavily used by the scientific

community, Python was chosen for the readability of its code (over R code, at least) and its gradual learning curve.

To use LattesLab, it is necessary to have all the Lattes CV files of interest stored in a folder and pass the folder name as a variable to the LattesLab main library. Then, the tool reads the CV files as downloaded from the Lattes Platform, parsing the XMLs and storing the data available on data frames.

To extract data from the Lattes CV, XPath Language - described in section 2.5.1 is used. One of the major advantages of the Lattes CV is its XML structure (which allows the extraction of semantic information from it) and the fact that the way its structure is described in an XML Schema Definition. By accessing the XML file through the XPath language one can extract the desired information and use it accordingly. In the case of this work, the extracted information is used to generate data frames that contain the data to be analyzed.

To obtain meaningful results from the analysis, it is desirable that the set of Lattes CVs share some characteristics. For example, to perform an analysis of the students and researchers of one institution, it is necessary to collect the Lattes CVs of members of that institution and analyze this specific set with LattesLab. Other examples of thematic Lattes CV sets are researchers of all institutions that share some CNPq classification (e.g. CNPq grantees), or researchers of a specific knowledge area.

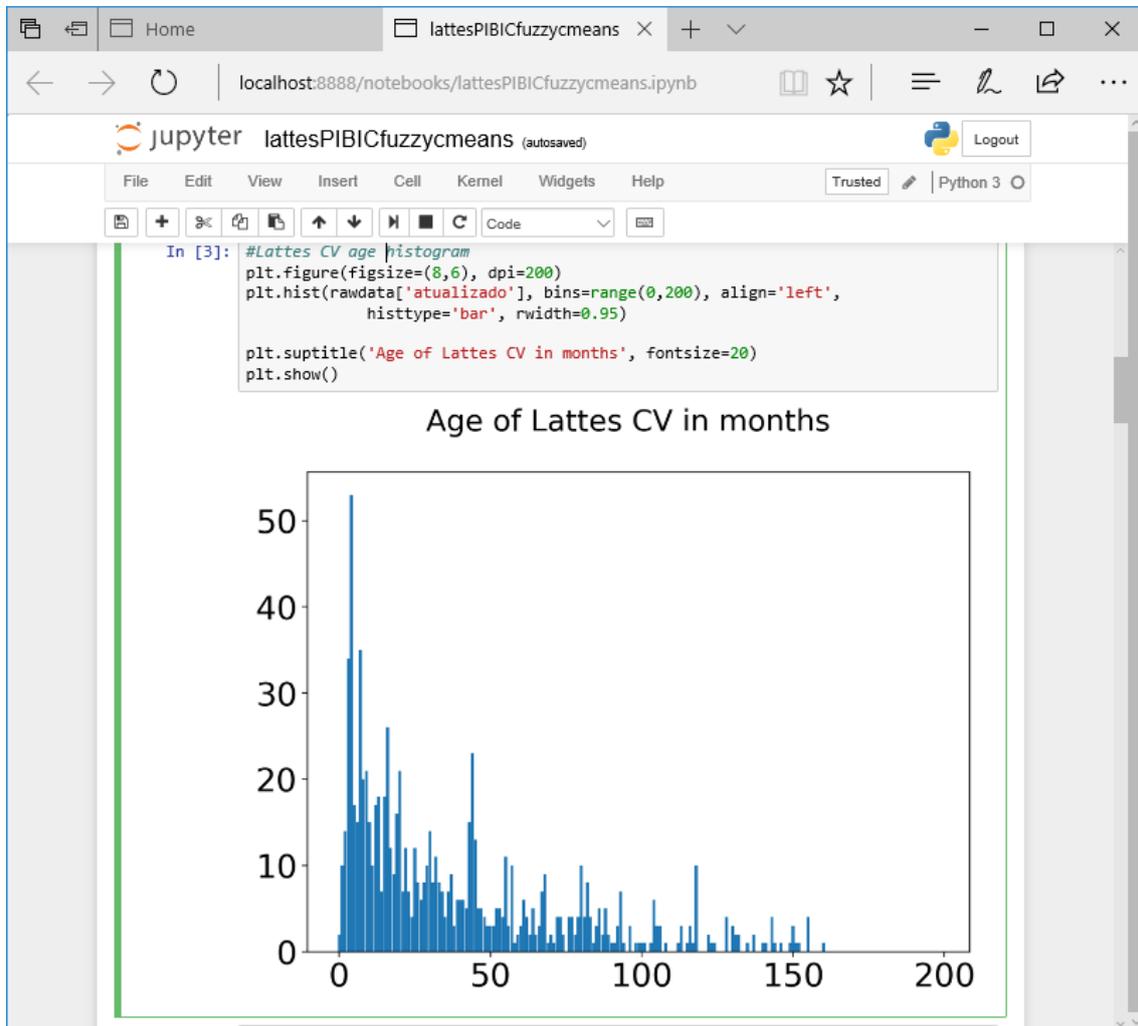
The LattesLab library is packed as a Python package, which simplifies its deployment. It can be downloaded and used in a standalone way, in this case the user must be able to at least install a Python IDE and then install and import the package.

Another deployment solution is to use LattesLab in a Jupyter notebook (KLUYVER et al., 2016). Jupyter notebooks are web applications that allow creation and sharing of documents that contain live code, equations, visualizations and explanatory text. The non-static parts of these documents are created by the execution of Python code.

Notebooks are a useful solution not only to facilitate access to the information, but also due to the fact that it runs in any operating system with a browser installed, and that it works not only with Python, but with R, Scala, Julia, and over 40 different programming languages.

Jupyter notebooks allow not only the deployment of code but, in the same document, formatted text (with different styles and allowing the use of hypertext, graphics, etc.)

Figure 3.3 - Example of a Jupyter Notebook window running the LattesLab tool.



SOURCE: Screen of a web browser running a Jupyter Notebook with the LattesLab library.

related to that code.

Therefore, by using a Jupyter notebook, it is possible to run LattesLab code, to perform analyses and visualizations, to explain the algorithm and to give instructions to the users, so that he or she doesn't have to know the tool beforehand to use it and can change parameters or modify analysis to suit specific needs. Figure 3.3 shows a Jupyter notebook running LattesLab.

When considering how to develop and deploy a tool for analysis of Lattes CV data,

the option to deploy a GUI-based application was evaluated, but not implemented. The reason to work on a tool with interactive lines of code is to give flexibility that such a tool requires. Consider a simple analysis that requires the selection of a date range on a set of CVs: a GUI must provide a widget to allow the input of an initial and end year, which is quite simple to implement and use, and a programming approach would require one or two lines of code to implement the same functionality. However, for more complex filters, e.g. to select a non-continuous range of dates, a GUI dialog would be more complex for the user (probably implemented as a list of check-boxes, one for each year) than one or two lines of code that filter a data set by a list of years.

A programming environment, while more complex, gives freedom to the user to implement filters, apply visual effects on graphics and use third-party tools, but the most important reason to avoid a GUI-based approach is the easiness of reproducibility: the chain of commands that provide an analysis from a data set can be expressed in code, which can be documented and read by users, while GUI-based applications would require gestures (clicks, scrolls, inputs) that must be preserved somehow to allow reproduction of the analysis.

3.4 Deployment

There are two ways to install and start using the LattesLab tool:

- Through a Python language IDE, and
- Through Jupyter notebooks.

These alternatives are detailed in the following sections:

3.4.1 Python IDE Deployment

The LattesLab tool is published as a Python library, and can be installed through the Python package manager *pip*, through the following code:

```
pip install LattesLab
```

It is also possible to install it by downloading LattesLab to a directory of your choice and use the setup script:

```
python setup.py install
```

The files can be downloaded from the official LattesLab repository in GitHub⁵.

Installation of the Python library is necessary regardless if one is going to write their own code or use the codes presented in the Jupyter Notebooks that are downloaded with the library.

The library is distributed according to the MIT License, which means that the code can be modified according to the terms of this license.

This distribution is aimed to the public that has some familiarity with the Python language, and that may want to use their skills to develop their own solutions to their problems.

3.4.2 Jupyter Notebook Deployment

The LattesLab library, on its installation, downloads a series of “standard” Jupyter notebooks. These notebooks contain different examples of applications of the LattesLab library, containing code and examples.

A list of these notebooks may be found in the “notebooks” folder of the installation, and online, in the LattesLab repository in GitHub⁶.

The distribution through Jupyter notebooks allows the user to run these notebooks and achieve immediate results. The notebooks also contain explanations of the code, guiding the user on how to use the LattesLab library and encouraging customization of the notebook code.

Having presented the LattesLab Python package, the next step is to exhibit, in chapter 4, a concise description of the Lattes CV database used for this analysis.

⁵<https://github.com/tcodingprojects/LattesLab>

⁶<https://github.com/tcodingprojects/LattesLab/tree/master/notebooks>

4 THE LATTES CV COLLECTION

In this chapter, the collection of Lattes CVs that is used throughout this dissertation is described. Since it is a collection of Lattes CVs from INPE alumni that have been part of the Scientific Initiation Program, this program is also presented.

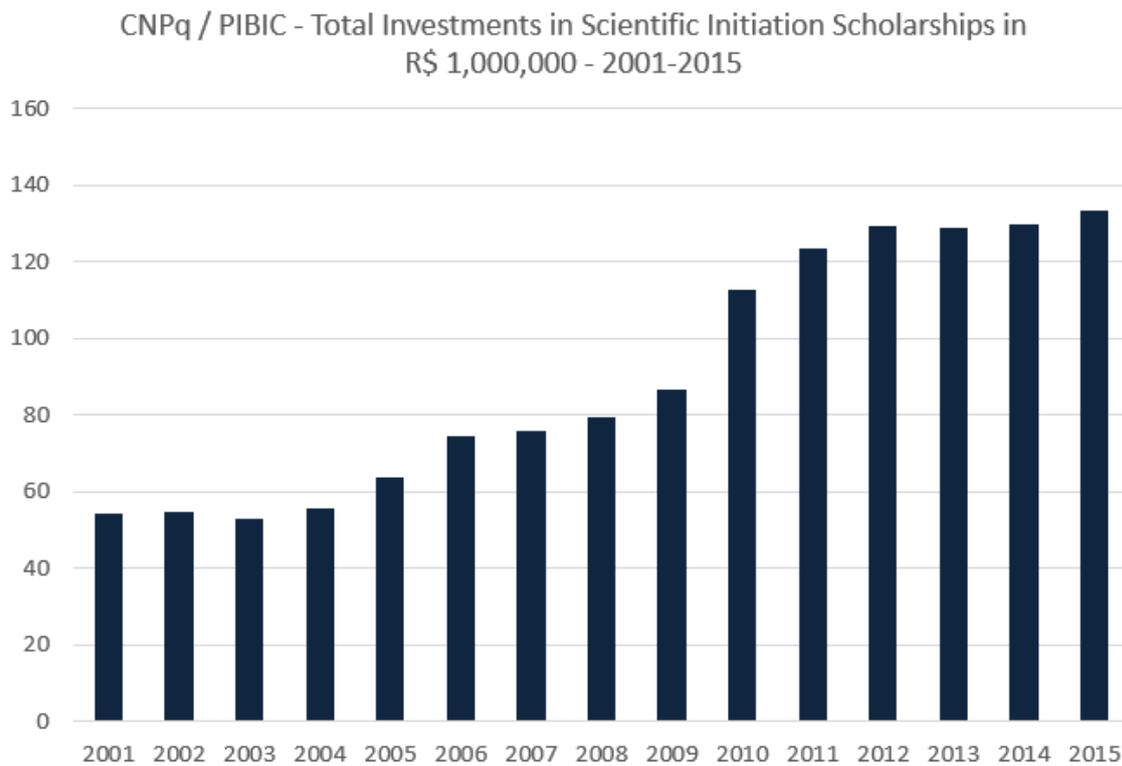
4.1 INPE's Scientific Initiation Program

INPE's role in the generation of knowledge and teaching of research skills to its students is achieved not only through its Postgraduate program, but also for its Scientific Initiation program.

Scientific Initiation programs have as target-audience undergraduate students with interest in learning and improving their research skills. These programs usually last 12 months, and have a monthly scholarship granted to the student.

In Brazil, the most noted Scientific Initiation programs are financed through CNPq, through a program named PIBIC *Programa Institucional de Bolsas de Iniciação Científica* (Scientific Initiation Scholarships Institutional Program) in English. The evolution of CNPq investments in Scientific Initiation scholarships is presented in Figure 4.1.

Figure 4.1 - Total Investments in Scientific Initiation Scholarships in BRL millions.



SOURCE: Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (2018a).

As stated in section 1.2, INPE has several post-graduation programs that promote research and graduate students in Masters and PhDs degrees. These programs also promote Scientific Initiation scholarships for undergraduate students INPE's.

Lattes CVs used in this work come from 906 INPE researchers that were part of the Scientific Initiation program. This collection's data is discussed in section 4.2.

4.2 Lattes CV Collection

For any thorough analysis of a data set, one question must be answered: why was this data chosen?

There are two possible answers to this question:

- Each individual component of the analyzed set share a single characteristic with each other. Therefore, an analysis of the data set generates information that can be associated to that specific feature that defines the data set.
- The individuals do not necessarily have any attribute in common. In this case, the analysis should be conducted to find some shared features between the individuals of the group.

The shared feature of the Lattes CV data set analyzed in this dissertation is that they all belong to researchers that were part of INPE's Scientific Initiation program.

According to (MITCHELL, 1969), social networks are defined as *"a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved"*.

Therefore, any Lattes CVs collection of researchers that share a given attribute can be classified as a social network.

The attribute which the Lattes CVs analyzed in this work share is that all of them are associated with INPE, because they were part of the Scientific Initiation program. However, other data sets could be part of a similar analysis, such as:

- Researchers that were part of other graduation or post-graduation programs.
- Researchers that were in either one of these programs in the 1990's.
- Researchers that possess a Bachelor's degree in Computer Science.

The data retrieved from these Lattes CVs generates a collection that is discussed in the next section.

4.3 Exploratory Data Analysis

According to (SEMATECH..., 2013), Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis which employs a variety of techniques (mostly graphical) to:

- maximize insight into a data set;

- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

EDA is usually one of the first Data Science approaches used to find out what information is contained inside the data. Through this initial analysis, it is possible to detect problems, inconsistencies, formulate and test hypothesis and explore the data, often making use of visual methods.

Through visualization, it is possible to comprehend information more quickly and effectively than through exploring piece by piece of the data.

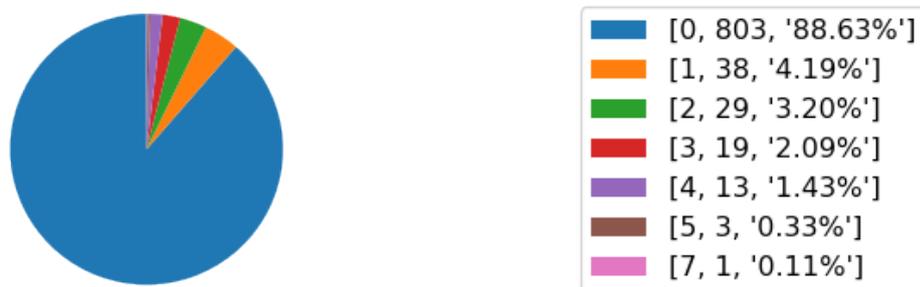
Since the data contained in the Lattes CV collection is composed of strings and numbers and there are "only" 906 CVs in the analyzed collection, two-dimensional graphics are sufficient to condense the information from the Lattes CV data and perform an EDA.

4.3.1 Scientific Initiation Scholarships per Student

The first visualization is that of the frequency of the Scientific Initiation (PIBIC) scholarships per student.

Figure 4.2 - Number of Years of Scientific Initiations per Researcher.

Scientific Initiation Freq



SOURCE: LattesLab library output.

Figure 4.2 shows that most of the researchers that declare to have taken part on a Scientific Initiation program, were part of this program for one year.

It is important to notice that even though this database is composed of Lattes CVs of students that have conducted at least one year of Scientific Initiation program, more than 88% of them have not declared that they were part of these programs. It is an indication of an important factor on the results of any Data Science experiment: the quality of the input data.

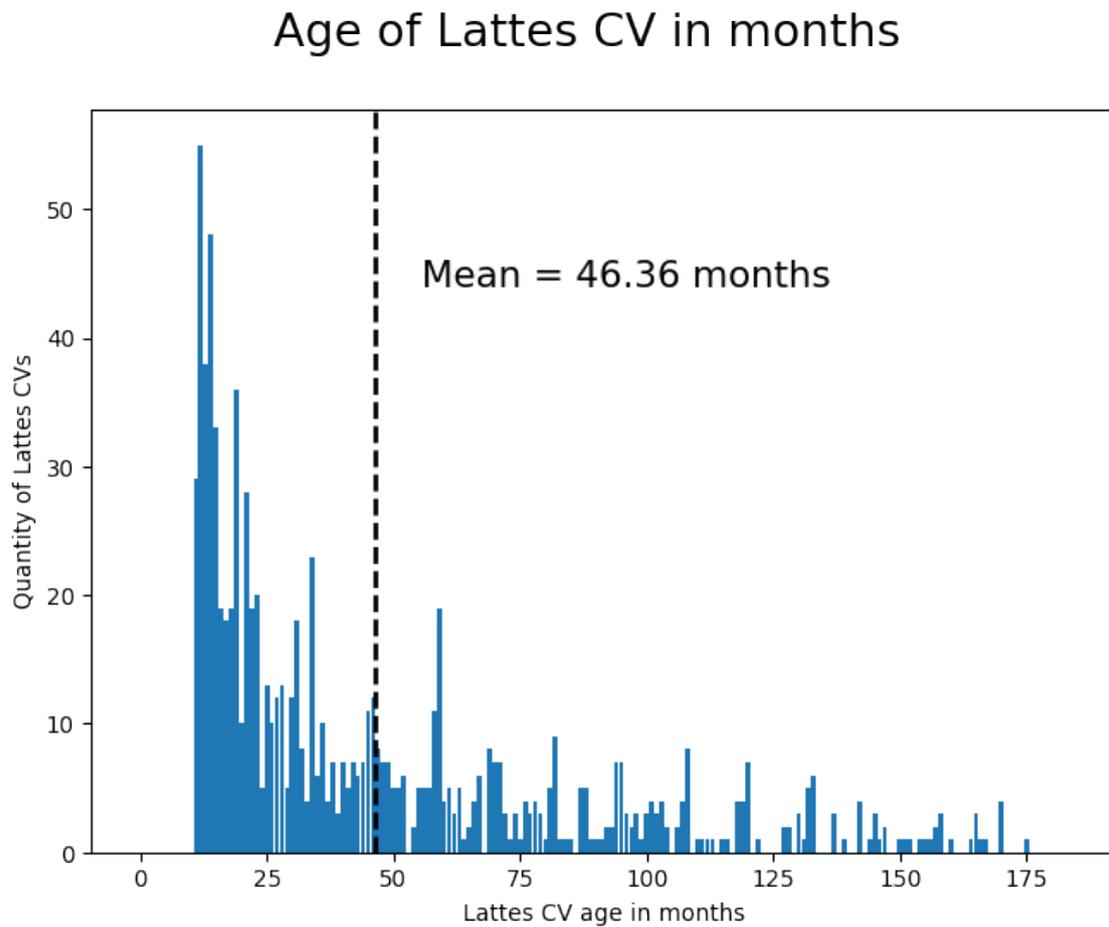
The LattesLab tool is dedicated to process the supplied data and extract results from it. It is possible that unexpected results are found. If the source of these errors is the interpretation of data, it can be found and fixed, but if the raw data source is incorrect, the approach is to acknowledge this incorrect data and, if desired, remove it from the data set.

However, there is a story behind each piece of data. It could be the case, for example, that a given researcher has not completed their Scientific Initiation program. This is something that cannot be interpreted by the Lattes CV data, nor by the results of LattesLab.

4.3.2 Lattes CV Last Update in months

Another relevant piece of information is when the Lattes CV was last updated. The histogram in Figure 4.3 shows the number of months from May 9th, 2018 (date in which the figure was generated) to the date each Lattes CV was updated.

Figure 4.3 - Number of Months Since Last Lattes CV Update.



SOURCE: LattesLab library output.

It is possible to see that there are Lattes CVs that haven't been updated for over 15 years. This information itself doesn't mean that the data contained in these CVs is wrong: the researcher may have not authored any publications, works, or concluded any graduation courses to add to their Lattes CV since the day the Lattes CV was last updated. However, the more recently updated the Lattes CV, the higher is the degree of confidence in the information it contains.

Therefore, it is possible to see that, with a mean "age" of 46 months, the owners of these Lattes CVs have not been updating their Lattes CVs frequently. It may be the case that most of them have simply not pursued the scientific career.

Figure 4.4 - Researchers Nationality

Nationality Frequency



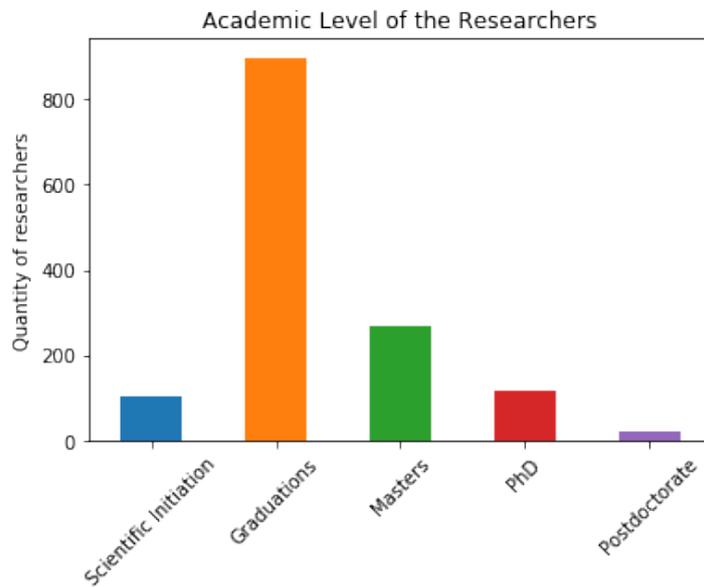
SOURCE: LattesLab library output.

4.3.3 Other Lattes CV Data

Another data that is listed in the Lattes CV is the nationality of the researchers. Since the Lattes CV collection analyzed contains CVs from researchers that took part in Scientific Initiation programs while they were pre-graduation students, it is expected that the majority of these students are Brazilian. Figure 4.4 confirms this hypothesis.

It is already known that these researchers were part of Scientific Initiation programs. But what about the continuity of their scientific career? Figure 4.5 shows how many of the researchers have continued with their studies.

Figure 4.5 - Academic Level of the Researchers



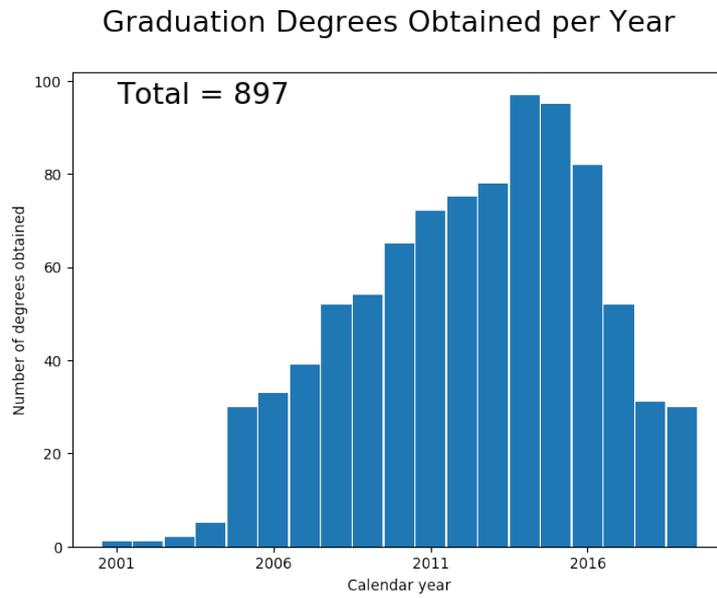
SOURCE: LattesLab library output.

It can be seen that most of the researchers have concluded their graduation. 266 of the researchers have concluded their Masters degree, while 116 have concluded their Doctorates and 22 are Post-Doctors.

Through the Lattes CVs, it is possible to retrieve the years in which the course (graduation, master's degree, etc.) has begun and the year it has ended. However, not always these years are supplied together, so that to infer the year of conclusion of a given course, a standard duration has been implemented. This duration is of four years for graduation and doctorate degrees and of two years for master's degrees and post-doctorate degrees.

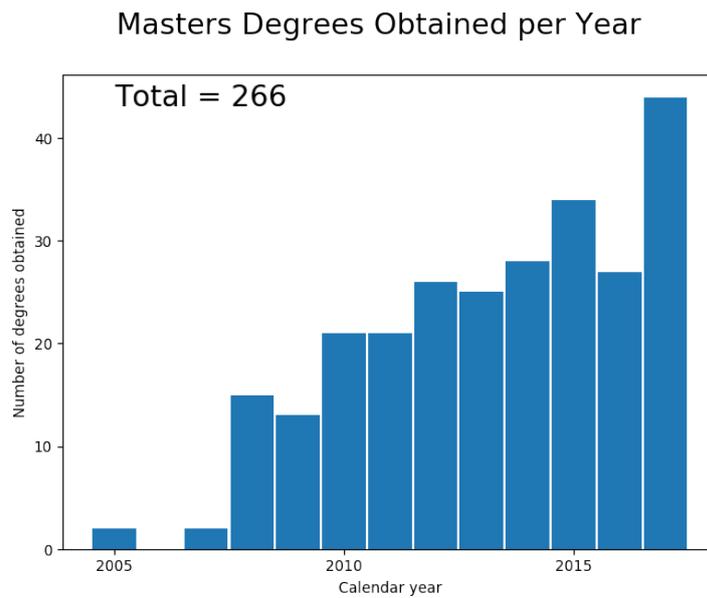
It is also possible to see how these degrees were obtained over time. Figures 4.6, 4.7, 4.8 and 4.9 show this evolution.

Figure 4.6 - Distribution of Graduation of Researchers per year.



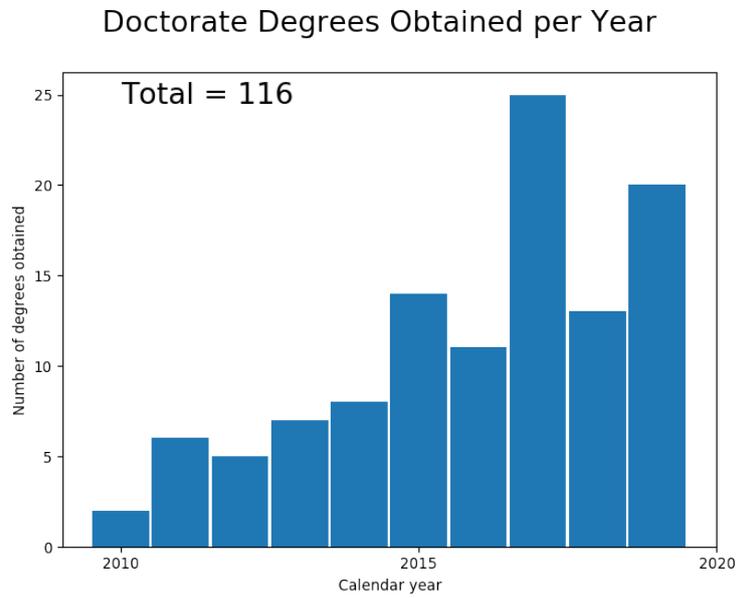
SOURCE: LattesLab library output.

Figure 4.7 - Distribution of Master's Degrees of Researchers per year.



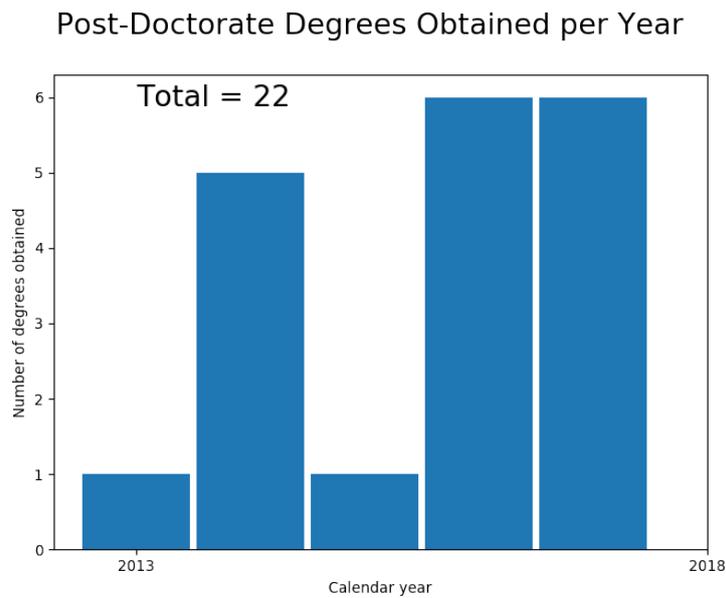
SOURCE: LattesLab library output.

Figure 4.8 - Distribution of Doctorate Degrees of Researchers per year.



SOURCE: LattesLab library output.

Figure 4.9 - Distribution of Post-Doctorate Degrees of Researchers per year.

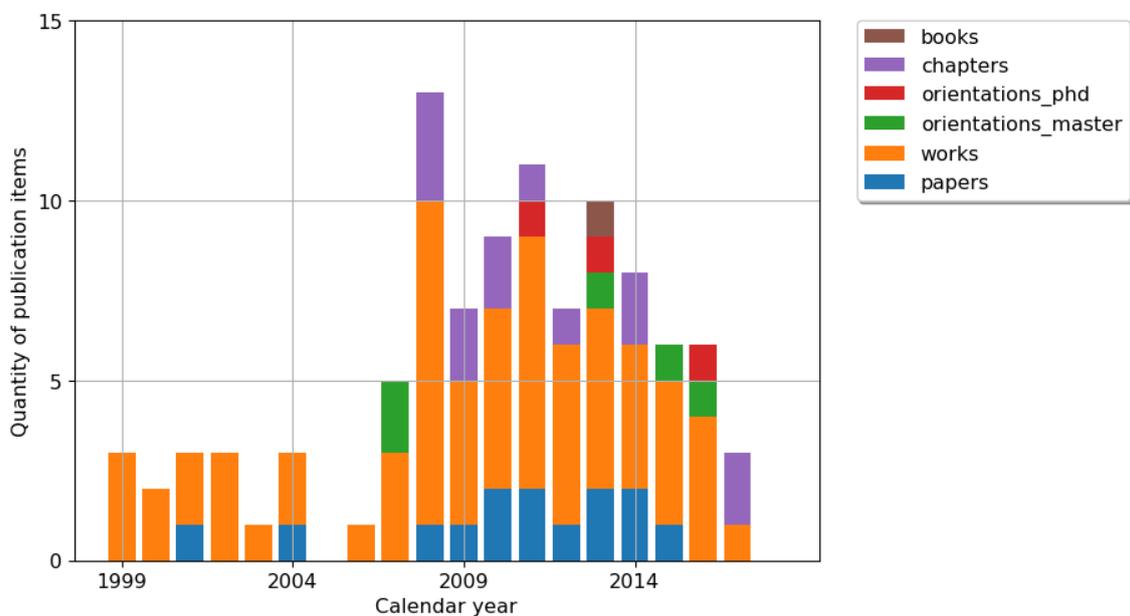


SOURCE: LattesLab library output.

Another analysis that can be performed using LattesLab is that of the history of production of a researcher. This algorithm searches the Lattes CV for the different types of works a given researcher has published per year and summarizes it in a bar chart.

In this analysis, it is possible to quantify the books and chapter of books the researcher has written, the number of orientations (for both Master’s and Doctorate degrees) as well as research projects and papers. Figure 4.10 shows an example of the results of this analysis for one of the researchers.

Figure 4.10 - Publication history of a researcher, obtained through Lattes CV Data.



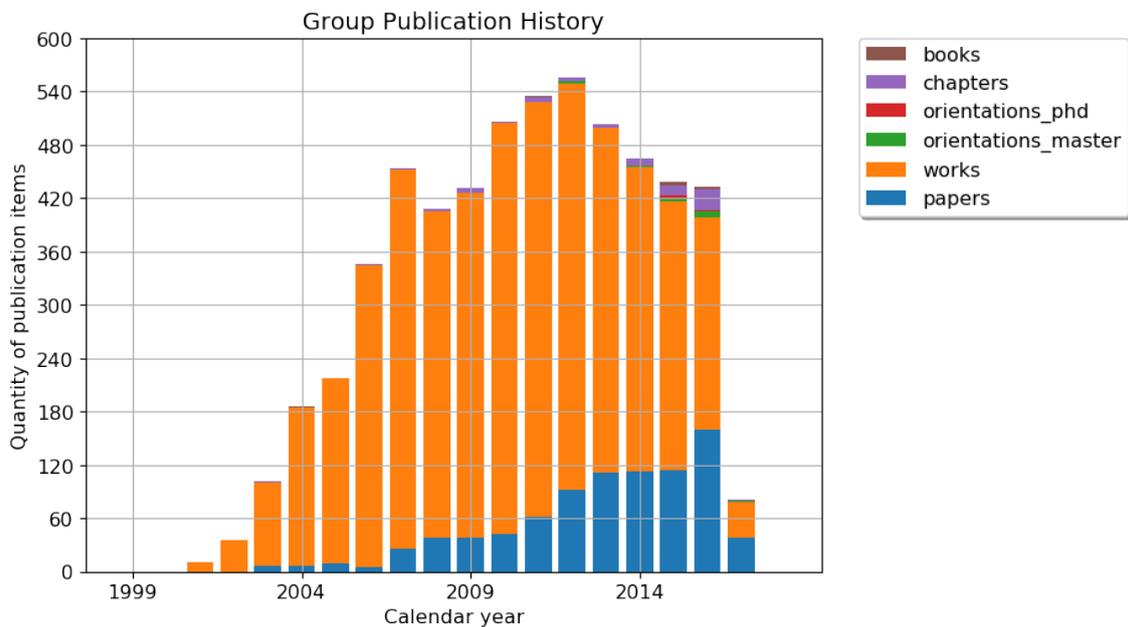
SOURCE: LattesLab library output.

Even though this analysis helps to identify the impact of a researcher’s work through numbers, this quantitative analysis shouldn’t be used exclusively, for there are nuances of the researcher’s productivity that this analysis do not capture. Not two papers have intrinsically the same relevance, for there may be different number of citations, different scientific impact depending on the subject of the paper and on its results.

This risk of oversimplification is present whenever an "index" is proposed or a data analysis is conducted. Summarization of data is useful, but the analyst must be aware that the raw data – in this case the Lattes CV – can say more than what is visible in a graphic such as the one in Figure 4.10.

The same analysis found in Figure 4.10 can be extended to a group of researchers, as it is shown in Figure 4.11 - which collects the Lattes CV data from the 906 scientific initiation researchers.

Figure 4.11 - Publication history of the group of researchers.



SOURCE: LattesLab library output.

4.3.4 Wordclouds

LattesLab is also capable of generating word clouds from the titles of the publications listed in the Lattes CVs of the researchers. LattesLab can also generate word clouds from the text contained in the Lattes CV summary section.

Word clouds are a direct and visually appealing visualization method for text. They are used as a means to provide a text overview by distilling its contents down to the

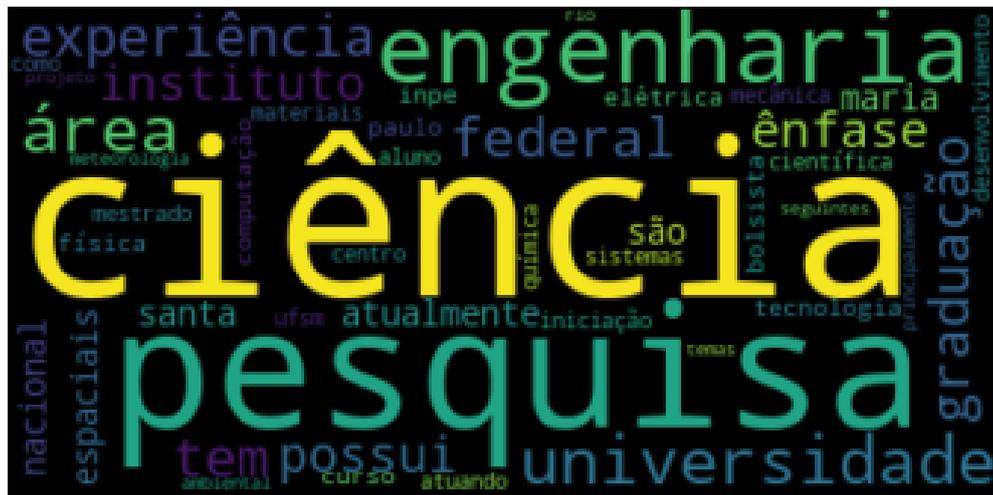
words that appear with highest frequency. Typically, this is done in a static way as pure text summarization (HEIMERL et al., 2014).

This summarization of the publication titles or the Lattes CV summary is useful for identification of keywords directly linked to the publications of a given researcher - or a group of researchers. This evaluation can be useful when a student is looking for an advisor, a researcher is looking for partnership with other researchers, and whenever a quick analysis of summaries or publication titles found in one or more Lattes CV must be performed.

To execute this word cloud processing, the Python library *wordcloud*¹ was used. This library contains code that generates a word cloud based on a list of words supplied to its functions.

Figure 4.12 contains the word cloud generated from the words contained in the Lattes CV summaries of all researchers. Figures 4.13 and 4.14 contains the word clouds generated from the titles of the publications of a given researcher and from the titles of the publications of all researchers, respectively.

Figure 4.12 - Word Cloud of the words in the Lattes CV summaries of all researchers.

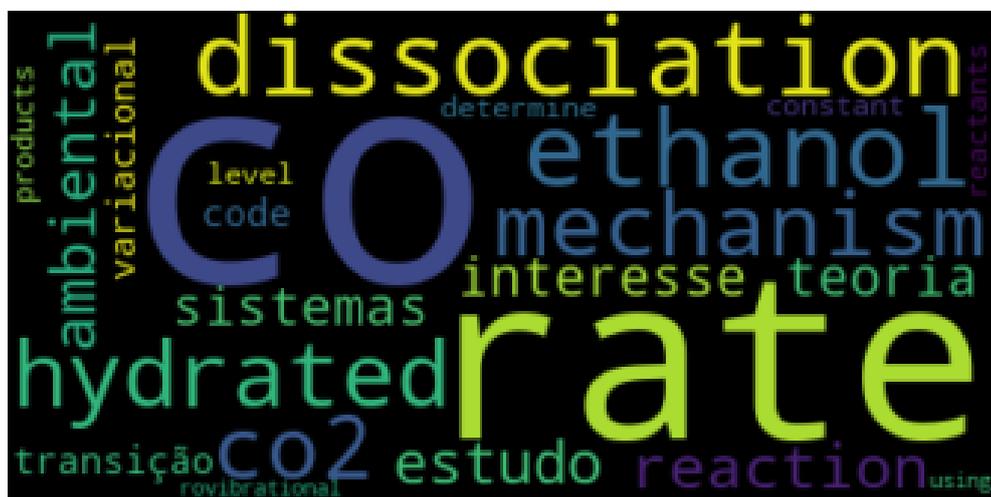


SOURCE: LattesLab library output.

¹https://github.com/amueller/word_cloud

Figure 4.13 - Word Cloud of the words in titles of publications of the Lattes CV of a given researcher.

WORDCLOUD:



SOURCE: LattesLab library output.

or group of researchers. Since there are generic words associated with any field of research - such as “science”, “study” or “analysis”, and that it is likely that these words appear with high frequency throughout a numerous quantity of Lattes CVs, it is expected that a word cloud generated by contents of several Lattes CVs contains a high percentage of these words - and therefore is less specific.

There are two ways to avoid this effect:

- Focusing the analysis on word clouds of individual researchers; or
- Removing generic words from the analysis. The *wordcloud* Python library allows a list of words to be removed from the analysis. This parameter is of extreme importance, since these analysis may easily be polluted by articles, prepositions and adverbs. It is also possible to add undesired words that may pollute the analysis to this list. In this implementation, the following words were removed from the analysis:

```
["quot", "of", "the", "in", "and", "on", "at", "for",  
"to", "by", "an", "with", "from", "com", "de", "em",  
"um", "uma", "do", "da", "dos", "das", "para", "no",  
"na", "nos", "nas", "por", "pelo", "pela"]
```

4.3.5 Graphs

The set of publications found in the Lattes CVs, when correctly described by the researcher, contains information about the other researchers with whom the Lattes CV owner has worked with. Through this characteristic of the Lattes CV, it is possible to unveil collaboration networks that are related to a given researcher or group of researchers.

According to (LIU et al., 2005), collaboration is able to promote research activity, productivity, and impact, and therefore is to be encouraged and supported by the means of research management and science policy.

In the case of the Lattes CV analysis, the connection between the individuals can be established for a series of relationships, such as:

- If they have collaborated in some kind of scientific work.
- If one has advised the other in their scientific career.

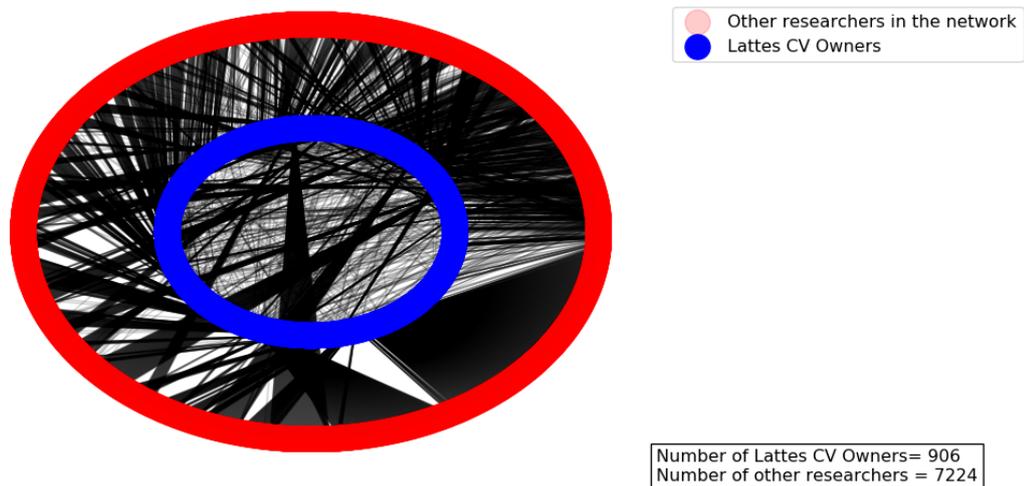
Even though researchers can share some characteristics – such as the school in which

they have concluded their graduation, or the course in which they have graduated – this doesn't translate automatically in the fact that these researchers share a personal or professional connection.

With LattesLab, it is possible to have a visual indication of what these Lattes CV Social Networks look like. To establish these networks, the Python package *networkx*² was used.

The network formed by these researchers, according to the Lattes CVs, is part of Figure 4.15.

Figure 4.15 - Collaboration Network of the Researchers.



SOURCE: LattesLab library output.

To illustrate the relationships between the researchers, the network is split in two circles; the inner blue circle is composed of the researchers which the Lattes CV is being analyzed. The outer red circle is composed of researchers which have collaborated with the researchers from the blue circle, but whose Lattes CV was not retrieved.

From Figure 4.15, it is possible to observe that:

²<https://networkx.github.io/>

- The network's blue circle is composed of 906 researchers. The whole network contains 8127 researchers.
- Some researchers have a larger number of collaborators. In the lower right part of the Figure 4.15 it is possible to see a part of the red circle converging to a single point of the blue circle.
- Some researchers have no connection at all, which translated in the fact that they have not declared any publications in their Lattes CVs.
- Even though the researchers only share the fact that they were part of INPE's Scientific Initiation program, there has been some collaboration between them. This collaboration may or may not be associated with their work in INPE. An specific analysis would make the conditions of this collaboration clear.

Due to the quantity of researchers that compose this network, it would have been impossible to have each node of the network identified individually in a single figure – even though researcher identification may be necessary to complement and extend the analysis. To allow this identification, the LattesLab algorithm produces a text file associating each researcher with their index in the network. With that, it is possible to retrieve from the program the information related to that researcher and their connections.

The results of this analysis also rely heavily on the quality of the Lattes CV data. Therefore, if a researcher misspells the name of a co-author, and this co-author appears in other parts of the analysis - with their own CV, or in another collaboration - that researcher is considered multiple times, when it's not really the case.

This situation may also occur when a researcher is cited in different ways throughout their bibliography. So even if their co-authors use the correct citation of one researcher, if they possess different types of citations, such as abbreviations of different parts of their name, they may be counted as more than one researcher.

Solutions for these problems are discussed in section 6.4.

Once the Lattes CV data has been presented, an application of this knowledge of the data is used in chapter 5 to try to classify the researchers in profiles using unsupervised learning algorithms.

5 QUANTITATIVE LATTES CV ANALYSIS

5.1 Introduction

Up to this point, an Exploratory Data Analysis over Lattes CVs data was conducted. Some statistics for researchers and groups of researchers were explored, and graphics were used to describe and analyze the data. However, there is another aspect of EDA to be explored in the Lattes CV analysis: clustering.

5.1.1 Clustering Algorithms

"Clustering" is a terminology used in Data Science to describe algorithms used to identify and group points of data of a given, larger set that have some shared characteristic. Through clustering algorithms, it is possible to identify properties inside the data and classify that data in subgroups that represent a given property.

Using classification algorithms, it is possible to reduce the quantity of elements – points, data, individuals – that require analysis to a few, representative classes of these elements – using a process called Dimensionality Reduction ([YAN et al., 2007](#)).

Clustering algorithms are applied to data that hasn't been categorized. If the data has already been labeled, the problem to be solved is to train an algorithm to learn how to categorize the data – through the labeled training set – and then apply this trained algorithm in unlabeled data. There is a performance factor associated, which is relative to the rate of correct labels the algorithm predicts. In clustering, there is no prior categorization of the data and, through some categorization algorithm, the data set is labeled automatically ([JAIN et al., 1999](#)).

5.1.2 Lattes CV Publication Data

Each Lattes CV contains the history of publications and scientific works in which the CV owner has been one of the authors. It is propose to use this history as a the data source with which the owners of the Lattes CVs are labeled.

To retrieve this data, each one of the Lattes CVs has been parsed and each scientific paper or research work has counted as one publication. These publications have been grouped per year and formed a data frame containing the number of publications of each researcher for the last twenty years.

What is intended to achieve with this analysis is to answer the following question: are there expected kinds of profiles of publications associated with the researchers?

It is possible to hypothesize that there are different classes of researchers, and some of these classes are evident: those that have been part of a Scientific Initiation program in college, those who started publishing only during their post-graduation studies, those who continued publishing after their doctorate degree and others. Therefore, if there are different classes of researchers, one should expect that the way these researchers publish is different.

To test this hypothesis, it could be possible to use the academic degree of each researcher as a label, and then use an algorithm to guess the academic degree of one researcher based on his or hers publication history. This approach, however, is limited to only a handful of known categories – namely, the researcher’s highest degree of education –, and does not present itself as a clustering problem, but as a supervised classification one. It also doesn’t address the fact that, for example, there may be more than one category of publication for PhD students.

Therefore, for each researcher, a vector of twenty components – with each of its components referring to the quantity of publications of that researcher in the associated year – has been created. In the moment of its retrieval from Lattes CVs, each component of the vector corresponds to a different year, from 2018 to 1999 – therefore, corresponding to twenty calendar-years.

The first element of these vectors can be chosen in three different ways:

- Using the original data, so that the analysis and its results are based on calendar-years.
- Switching the first production value to be the first non-zero value of the productivity list. The last components of the list are substituted by zeros. In this case, the analysis is based on what is the researcher scientific production from the moment he or she starts publishing.
- Switch the first production year to be first year of Scientific Initiation scholarship. The analysis in this case is similar to the previous one, but in this case, the researcher may not necessarily have published in the same year he or she started the Scientific Initiation program. This is only a reasonable option in Scientific Initiation-based analysis.

The Lattes CV data gathered for this work belongs to participants of INPE’s Scientific Initiation program. It is known beforehand that all researchers whose Lattes

CV are used were part of a Scientific Initiation program. Therefore, the third option – i.e., to have the first year of each researcher to be the year in which they started their Scientific Initiation studies – is used.

5.2 Clustering Algorithms

To classify and group the 906 researchers of this work in different categories, two clustering algorithms have been used and are explained in sections 5.2.1 and 5.2.2: the Fuzzy C-Means and the Self-Organizing Map algorithms.

5.2.1 Fuzzy C-Means Algorithm

One characteristic of many clustering algorithms is that they associate each entry of the data set with one of the possible clusters, and may ignore the fact that a piece of data may share similarity with more than one of the clusters. These algorithms are called hard clustering algorithms (BEZDEK et al., 1984).

However, there is a class of clustering algorithms which consider the possibility that a piece of data can be associated with more than one data cluster. This clustering method is called fuzzy clustering (or soft clustering).

The Fuzzy C-Means, introduced by (DUNN, 1973) is one of these soft clustering techniques. It is similar to the K-Means (HARTIGAN; WONG, 1979) hard clustering algorithm: a number of clusters is provided, and each of these clusters is associated randomly to each one of the data points. The position of the clusters and the points associated to them are changed with the objective of diminishing the distance – as defined in the algorithm – between the data points to the centroid of its assigned cluster.

In the specific case of Fuzzy C-Means algorithm, it generates a partition matrix $W = w_{ij} \in [0, 1]$ – which dimensions are the number of data points and the number of clusters – in which each element w_{ij} refers to the degree the data point x_i relates to the cluster c_j .

The Fuzzy C-Means algorithm equals the K Means algorithm when for each row of the matrix W defined above, there is only one $w_{ij} = 1$ and all other $w_{ij} = 0$ – so that each data point is related to a single cluster.

5.2.2 Self-Organizing Map (SOM) Algorithm

Another widely used clusterization algorithm is the Self-Organized Map – also known as the Kohonen Map. It was introduced by the Finnish researcher Teuvo Kohonen (KOHONEN, 1982) in 1982, based on the fact that arrays of one or two dimensions can be used to generalize the topological projections of data, through a process of competitive learning, in which the distance of a data point is calculated for all of the cells, and the matching cells are adjusted to better match the inputs associated with their data points.

The Self-Organizing Maps can be used in a series of applications, such as analysis of financial stability, the fault diagnosis, the creation of well-composed heterogeneous teams and the application of the atmospheric sciences (JOHNSSON, 2012).

The algorithm works according to the following steps:

- a) Randomly select the grid's neurons positions.
- b) Choose one data point.
- c) Calculate the distance from that data point to each neuron. Find the neuron that has the minimum value of distance. This neuron is called the Best Matching Unit (BMU).
- d) Change the BMU coordinates so that it is closer to that data point. The changes in coordinates are ruled by a learning rate. This learning rate decreases after each iteration, to allow convergence of the algorithm.
- e) Move other neurons closer to the BMU's towards that data point as well, with neurons distant to the BMU moving less. These neighbors neurons are identified using a radius around the BMU. This radius' distance must also decrease after each iteration, to allow convergence.
- f) Update the learning rate and BMU radius, and repeat the algorithm.
- g) Perform the previous steps again until a maximum number of iterations is reached, or the positions of the neurons become stable.

5.3 Quantitative Results

5.3.1 Fuzzy C-Means Results

The Fuzzy C-Means algorithm was evaluated for a number of clusters that vary from three to twenty clusters. The performance of the error as a function of the clusters is presented in Table 5.1.

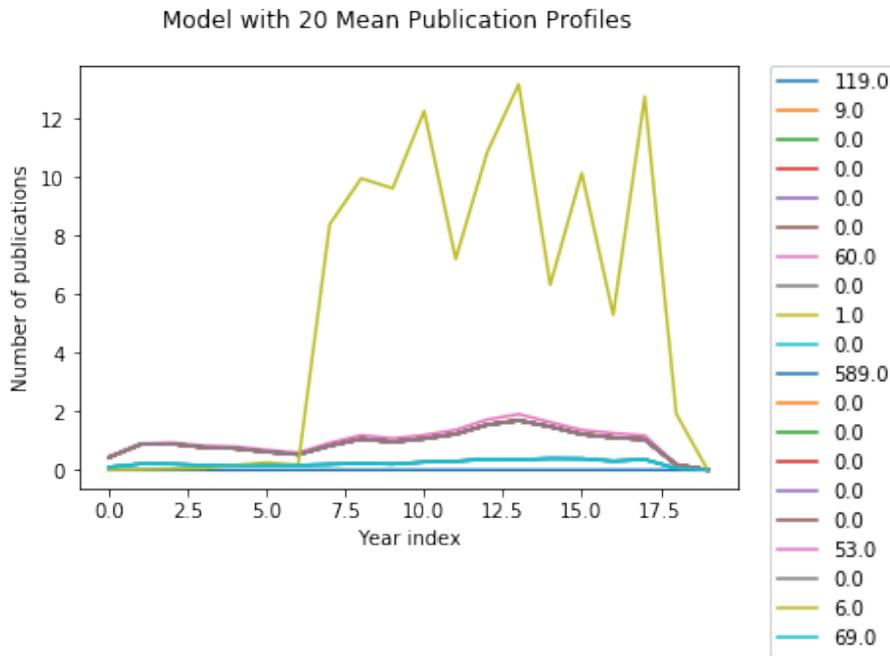
Table 5.1 - Error as a function of the number of clusters.

Number of Clusters	Error
3	0.76998
4	0.72345
5	0.69255
6	0.67054
7	0.65365
8	0.63978
9	0.63102
10	0.62329
11	0.61571
12	0.61050
13	0.60603
14	0.60255
15	0.59801
16	0.59543
17	0.59316
18	0.59112
19	0.58824
20	0.58745

As expected, the error reduces as the quantity of clusters raises. This behavior is expected since with an extra cluster, it is possible to better fit a piece of data with some residual error. However, raising the number of clusters can lead to an over-fitting of the model, i.e., having a model that only produces acceptable results for the specific set of data that was used to generate it, and do not capture the general behavior of the phenomenon observed.

It is possible to verify the fitness of the twenty-cluster model by observing the coordinates of its centroids. The legend of the Figure 5.1 shows the quantity of data points associated with that centroid.

Figure 5.1 - Centroid coordinates: twenty-cluster model.



SOURCE: LattesLab library output.

A few observations can be made about the results in Figure 5.1:

- Even though there are twenty centroids represented, there are only a couple of topological behaviors of interest: one that shows a big raise in the number of publications after the sixth year, a few profiles that fluctuate between one and two publications through the whole model, with peaks close to year thirteen, and two close to zero publications.
- Through the numbers in the legends, it is possible to see that only five centroids have over 40 data points associated with them. Also, eleven of the twenty centroids have three or less points of data associated with them.

Through these characteristics, it is possible to state that the model contains centroids that do not capture behaviors that apply to a significant part of the data. The model also contains centroids that present redundant topologies. Therefore, the conclusion is that the twenty-centroid model is over-fitted.

Also, since one obvious researcher behavior was detected, the data points associated with it can be detected and eliminated. This behavior is that of the researchers that have published zero scientific papers. Having observed this behavior, it is possible to eliminate from the data points the researchers with zero publications and make the algorithm focus in detecting other less-expected behaviors.

The performance of the algorithm after removal of the zero-publication points as a function of the clusters is shown in Table 5.2.

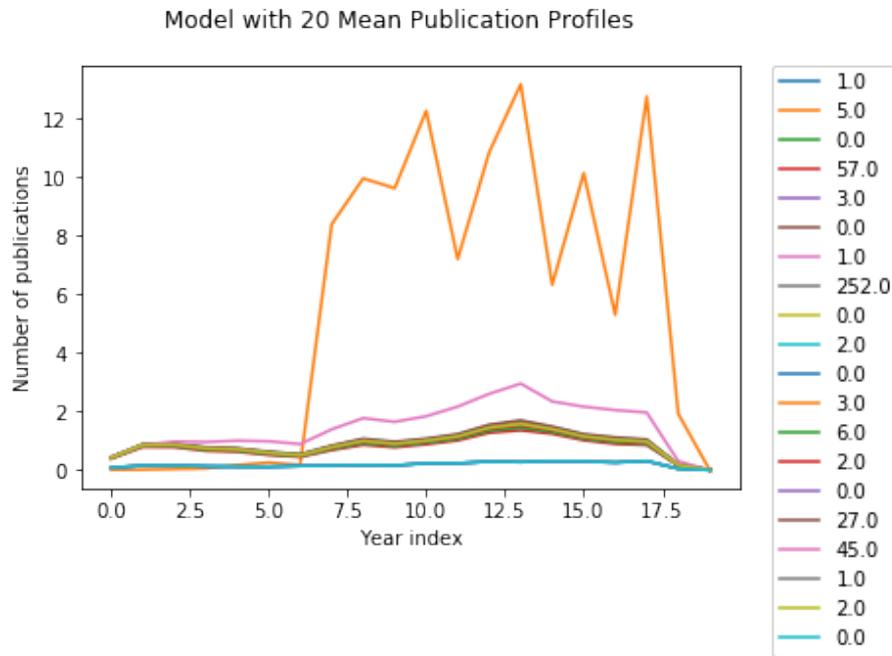
Table 5.2 - Error as a function of the number of clusters for the second experiment.

Number of clusters	Error
3	0.48400
4	0.38093
5	0.31432
6	0.26731
7	0.21405
8	0.19295
9	0.17760
10	0.16247
11	0.14106
12	0.13183
13	0.12293
14	0.11575
15	0.10410
16	0.09896
17	0.09028
18	0.09002
19	0.08595
20	0.07960

It is possible to observe that the error produced by the models after the removal of the zero-data points was reduced. The minimum error obtained by the previous model was 0.58745 – refer to Table 5.1 –, which is higher than the highest error obtained in the second experiment of the algorithm – that is 0.48400, according to the data in Table 5.2.

The twenty-cluster model for this new experiment is shown in Figure 5.2. Once again, it is expected that this model is over-fitted for the data with which it has been trained with.

Figure 5.2 - Centroid coordinates of the second experiment: twenty-cluster model.

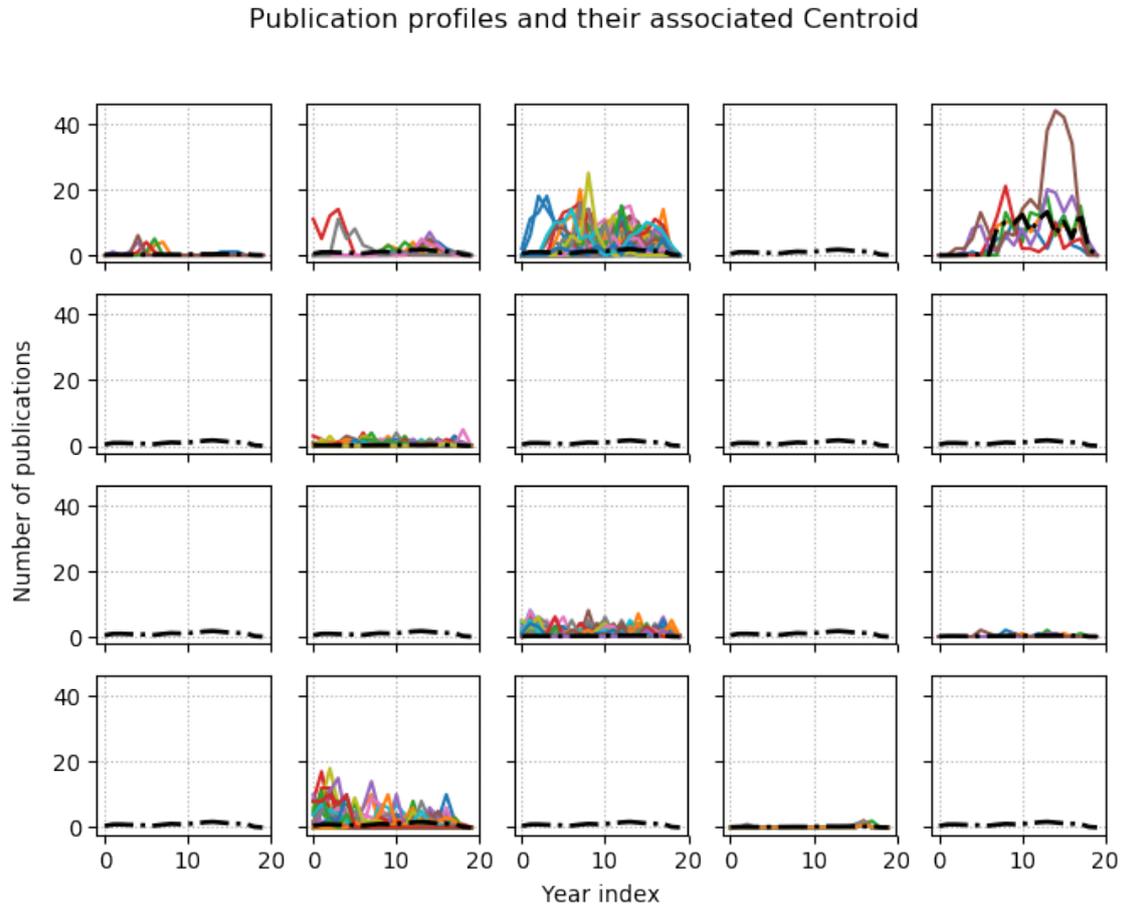


SOURCE: LattesLab library output.

This model displays three behaviours that were also observed in the previous experiment. This time, however, only four of the clusters have more than 40 data points associated with it, and the remaining clusters have six or less points combined to them. The number of clusters with zero associations is ten, half of the number of clusters of the model.

This association – or lack of association, depending on the centroid – can be observed in Figure 5.3, which shows each centroid and its data points.

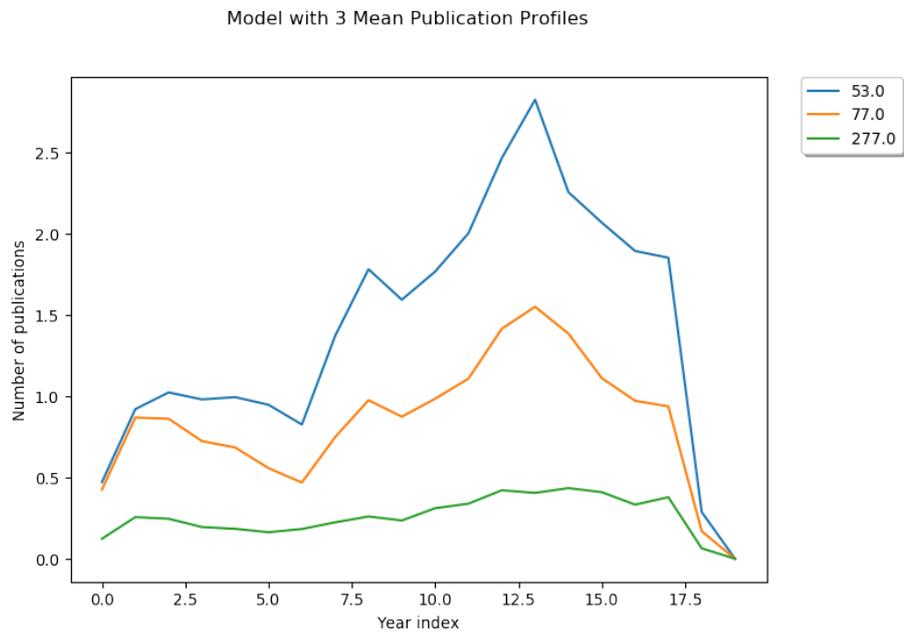
Figure 5.3 - Centroid coordinates of the second experiment and its associated data points: twenty-clusters model.



SOURCE: LattesLab library output.

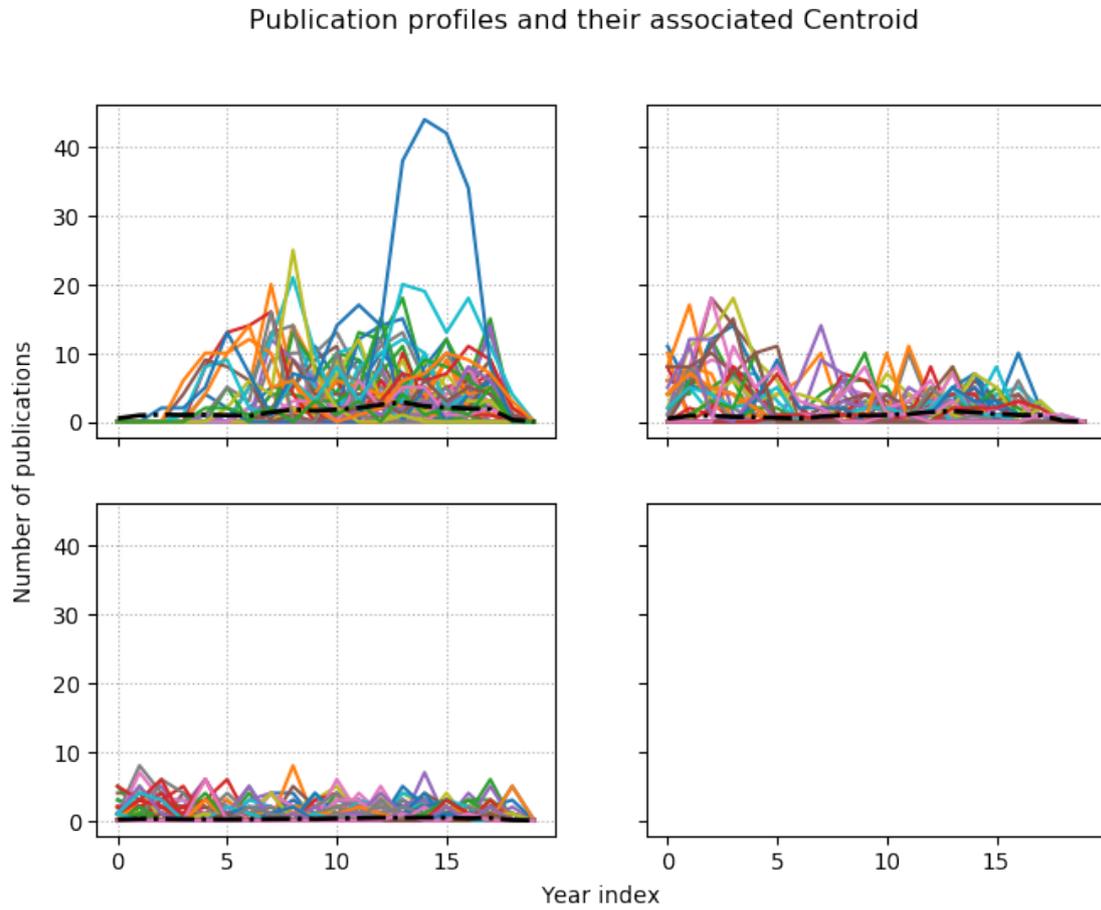
Since only three different topologies of publication's evolution were observed in this experiment, the same results are repeated for the model with three centroids. The results are shown in Figures 5.4 and 5.5.

Figure 5.4 - Centroid coordinates of the second experiment: three-clusters model.



SOURCE: LattesLab library output.

Figure 5.5 - Centroid coordinates of the second experiment and its associated data points: three-clusters model.



SOURCE: LattesLab library output.

It is possible to verify that the three-centroid model retains the capability of identifying the mean topologies of the data, without specializing in small portions of the data set, or repeating patterns throughout more than one centroid.

5.3.2 Self-Organizing Map (SOM) Results

The results obtained using the Self-Organizing Map model to identify mean publication models are presented in this section. For the SOM implementation, square matrices of neurons have been used. The data entries with zero publications have also been eliminated from the analysis, since it is an already acknowledged result.

The dimensions of these neuron matrices that were evaluated are: 3x3, 6x6, 9x9, 12x12 and 20x20. The error associated with each of these models is presented in Table 5.3.

Table 5.3 - Error as a function of SOM neuron matrix dimension.

Dimension of neuron matrix	SOM error
3x3	5.79215
6x6	5.74171
9x9	5.72291
12x12	5.72430
20x20	5.72075

It is possible to see that, as the dimension of the matrices increases, the error diminishes. In this SOM model – as in the Fuzzy C-Means – the analysis is also prone to over-fitting. It is possible to verify the fitness of the model by verifying the number of data points associated to each neuron. Figures 5.6 and 5.7 show these results, in numbers and graphically.

Figure 5.6 - Number of data points associated with each neuron of the 20 x 20 SOM model.

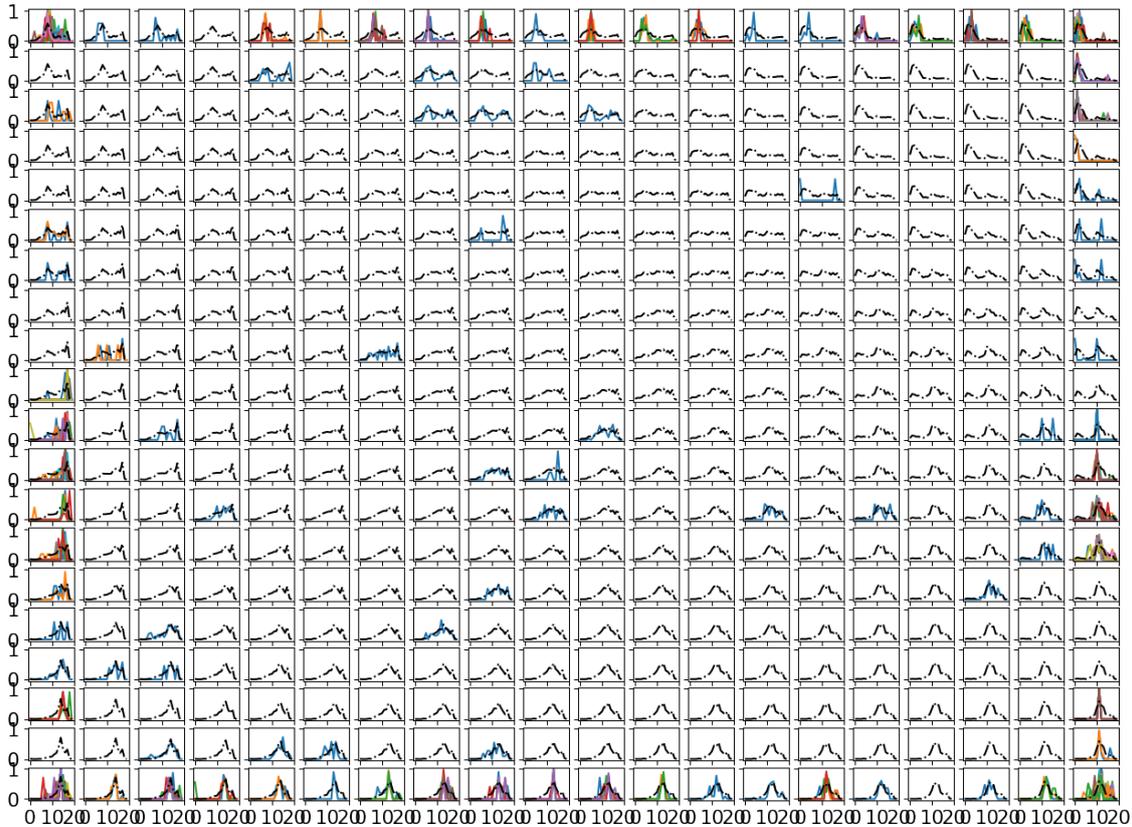
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
0	37	1	1	0	4	2	6	5	4	1	4	3	4	1	1	5	3	6	3	24	
1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	5
2	2	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	8
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
5	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	15	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	11
11	16	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	6
12	4	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	6	6
13	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9
14	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
15	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
18	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
19	35	2	5	4	2	1	3	6	5	5	5	3	1	1	4	1	0	1	3	33	

SOURCE: LattesLab and MiniSom libraries output.

The following observations can be made through the contents of Table 5.3 and of Figures 5.6 and 5.7:

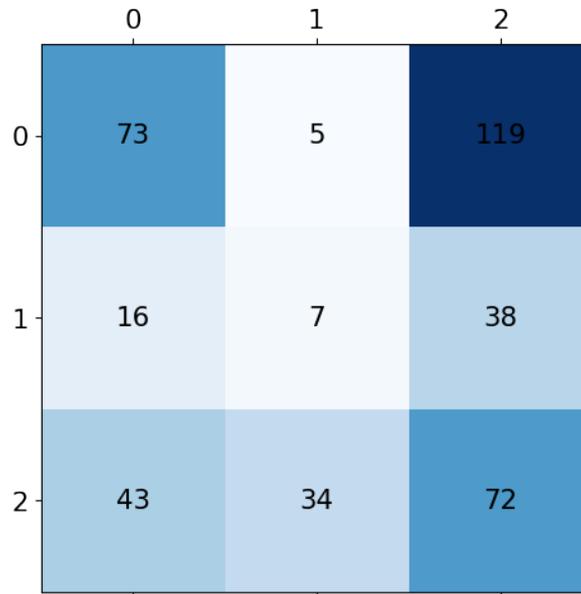
- Even though the error keeps diminishing as the number of neurons increases, the changes are not significant - specially for neuron matrices higher than 9x9, in which the error raises at 12x12, and is lowered in the 20x20 model. The change occurs only in the 10^{-3} digits.
- Both Figures 5.6 and 5.7 show that the majority of neurons are empty, having no data points associated with them. It also is observable that mostly the neurons on the boundaries have data points associated with them, and only nine neurons – of all 400 – have more than ten data points associated.

Figure 5.7 - Distribution of data points through each neuron of the 20 x 20 SOM model.



SOURCE: LattesLab and MiniSom libraries output.

Figure 5.8 - Number of data points associated with each neuron of the 3 x 3 SOM model.



SOURCE: LattesLab and MiniSom libraries output.

Having observed those behaviors, the model with the 3x3 neuron matrix is compared with the 20x20. Its results are found in Figures 5.8 and 5.9:

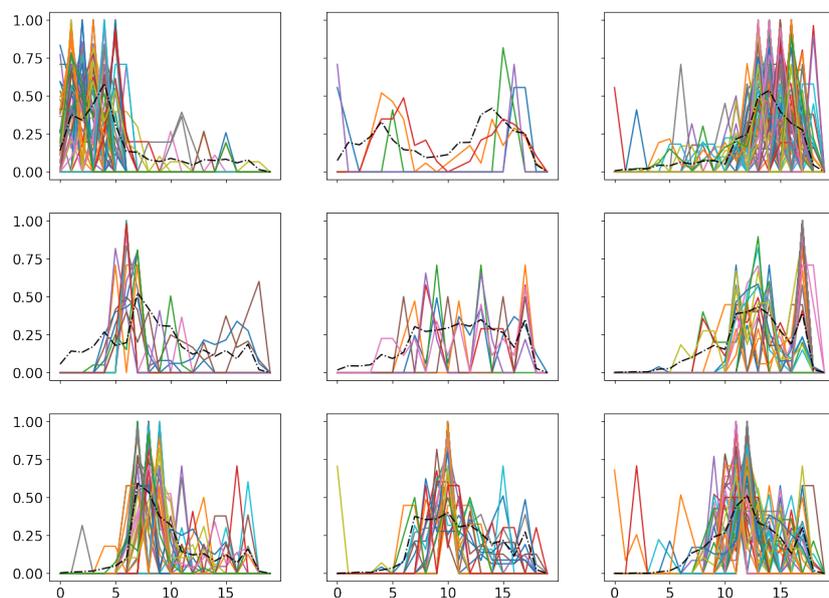
It is possible to verify that the neurons have more data points representing them than in the previous model. It is expected since there are less neurons to spread the data points with.

Also, in this 3x3 model, it is observable the influence of the neighbor neurons and how neighbors are similar to each other. The first column neurons all start with the last third part close to zero; moving to the right of the figure, that last third starts growing and the first and second third of the neurons diminishes. Also, the first row of neurons have the highest values in the first third of the neurons, the third row has the lowest, and the second row has values between the first and third rows.

5.4 Discussion

What do these results say about the existence of one or more mean publication profile of these PIBIC researchers? Both models used to analyze the publication data of these PIBIC researchers have provided the following insights:

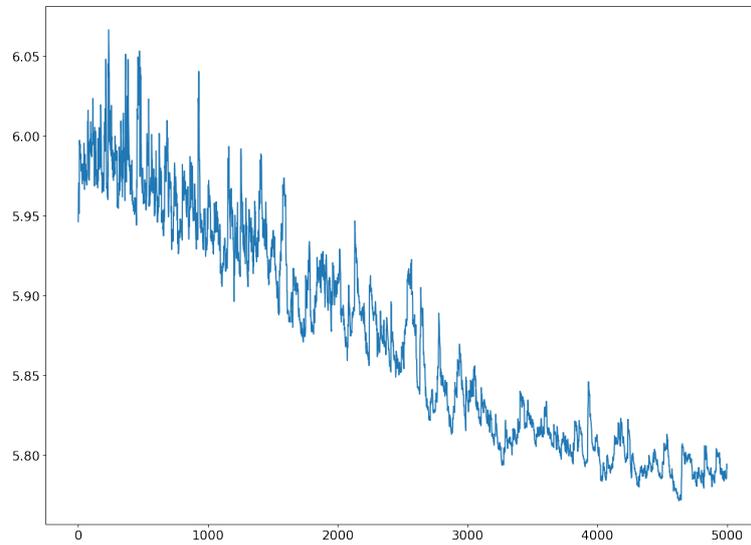
Figure 5.9 - Distribution of data points through each neuron of the 3 x 3 SOM model.



SOURCE: LattesLab and MiniSom libraries output.

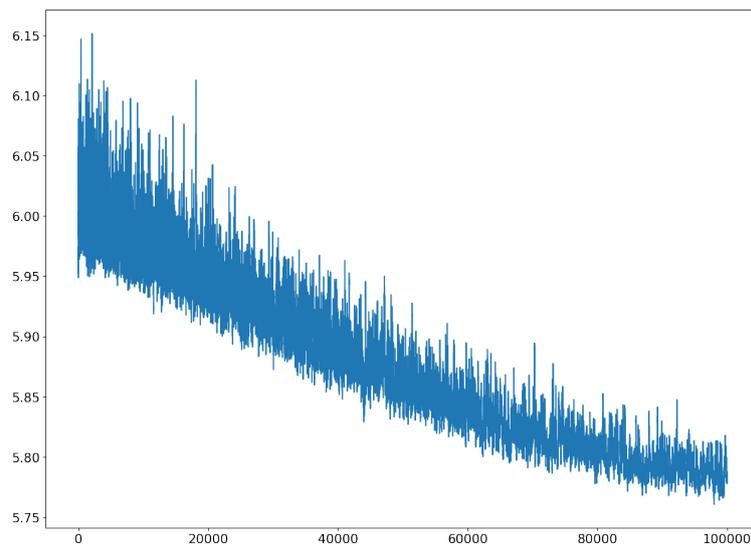
- The Self-Organizing Map produces a higher error than the Fuzzy C-Means – even for structures with higher number of neurons. Even when the number of training epochs is raised, the order of magnitude of the error does not change. This behavior is observed in Figures 5.10 and 5.11, where the number of training epochs was increased.
- These "mean-profiles" also contain non-integer components, which is not physically representative – even if one researcher is a co-author, the work where they are co-authors count as one.
- The assumption of twenty years of publications also influences the model and its results. The majority of the researchers whose data has been analyzed doesn't have a research career that span through twenty years. The assumption of twenty years of publications exists so that every researcher that started or ended a Scientific Initiation program in this time interval has their data analyzed. However, this time interval doesn't have to be also a range that influences the algorithm's capability of analyzing Lattes CVs of researchers that have indeed published for twenty years. Other ways to analyze this data are proposed in section 6.4,

Figure 5.10 - Evolution of SOM error for 3 x 3 neuron models and 5000 training epochs.



SOURCE: LattesLab and MiniSom libraries output.

Figure 5.11 - Evolution of error for 3 x 3 neuron models and 100000 training epochs.



SOURCE: LattesLab and MiniSom libraries output.

- Not only the scientific career for most Scientific Initiation program students doesn't last for twenty years, the data used in this experiment shows that most of the students failed to produce a scientific work during their participation of the program. These problems are revisited and possible solutions are discussed in section 6.4.

With the results presented and the discussion about how to achieve different – and better – results started, it is time to move to the conclusion of this work.

6 CONCLUSION AND FUTURE WORKS

In this chapter, the contributions of this work are summarized in section 6.1, section 6.2 brings a conclusion to this work, section 6.3 describes the scientific paper that has been published while this work was under development and section 6.4 proposes improvements and different ways to implement the LattesLab library, and to expand its capabilities.

6.1 Contributions

In this work, the following contributions were made:

- Chapter 1 presented the state of research, with special attention given to the Brazilian case, and proposed how Lattes CVs may be used as a means to quantify research and research quality.
- Chapter 2 described the Lattes CV environment and its evolution. It highlighted the fact that, being an open platform, it is at the same time rich in information and relaxed in the reliability of the data presented – what can be a complication in research works that use its data.
- In Chapter 3, a review of computational tools that were published and presented has been conducted. Several tools, their advantages and idiosyncrasies have been analyzed, and they have also been contextualized with the current state of the Lattes Platform. Requirements for a new tool were discussed and the LattesLab tool was presented.
- In Chapter 4, the LattesLab tool was used to analyze the structure of data inside a Lattes CVs. Exploratory Data Analysis (EDA) was conducted and visualization of Lattes CV data – in form of word clouds generated through Lattes CV summaries and through publication titles – was performed. Also the relationships of the researchers were identified and explored through graph analysis. The results presented in this chapter are convenient for users looking to summarize the information presented in Lattes CVs and to use this information in different ways, such as verifying the presence of inconsistencies, generating reports to funding agencies and boards, among other uses. In possession of a Lattes CVs set, word clouds generated with their data may be used to find areas of interest and help to find collaborators for a given project. Also, students may use these word clouds to

identify areas of research of prospect advisers, helping to make the process of finding an adviser easier. Finally, graphs may be used to identify social networks inside research departments and other environments, and then use this information to make a specified analysis of the relationship between these researchers.

- Chapter 5 applied LattesLab capabilities in an artificial intelligence scenario. For that, the following question was proposed: once the publication history of a group of researchers is retrieved, is it possible to identify mean publication profiles for the researchers of this group? Two clusterization algorithms – Fuzzy C-Means and Self-Organizing Maps – were presented and the results were also discussed in the chapter. The objective of this chapter is to demonstrate that the LattesLab tool is scalable and can be used to solve Data Science problems that are associated with Lattes CV data.

6.2 Conclusion

The main objective of this research is to propose a framework to retrieve, analyze and summarize results from a set of Lattes CVs. Due to the current state of the Lattes CV system – where automatic download of the Lattes CVs is hindered by CAPTCHA tests –, the proposed framework has to consider that users of the tool have already downloaded the Lattes CVs of interest which are subject of analysis.

This work also discussed alternative ways of getting in possession of Lattes CV files – such as the administrator of a given University, who could access and curate the Lattes CVs of the researchers from that university. The LattesLab library could have been developed for that target audience, but it is the understanding of the authors that it would restrict the use of the framework to a very specific audience. Therefore, the premise that the Lattes CV files have to be already downloaded before using LattesLab was adopted.

The authors proposed that the work with LattesLab is done through Jupyter notebooks – or at least starts with the code presented and explained in these notebooks. The Jupyter notebooks are didactic technical documents that contain not only programming code, but also explanations of what the code does.

It is also possible to acquire the LattesLab library in any Python IDE through the command:

```
pip install LattesLab
```

Through the IDE, the user can use the code present in the Jupyter notebooks, and make several other customizations, even in the code of the library itself.

6.3 Publications

During the development of this research work the following paper was published:

- *Data Science Approach to Analysis of Lattes CV Data* (SANTANA; SANTOS, 2017): in this paper, the concept of a Data Science approach to Lattes CV Data analysis was first proposed. There, the authors developed a prototype of the LattesLab library and achieved some initial results that were the basis of the work found in this dissertation.

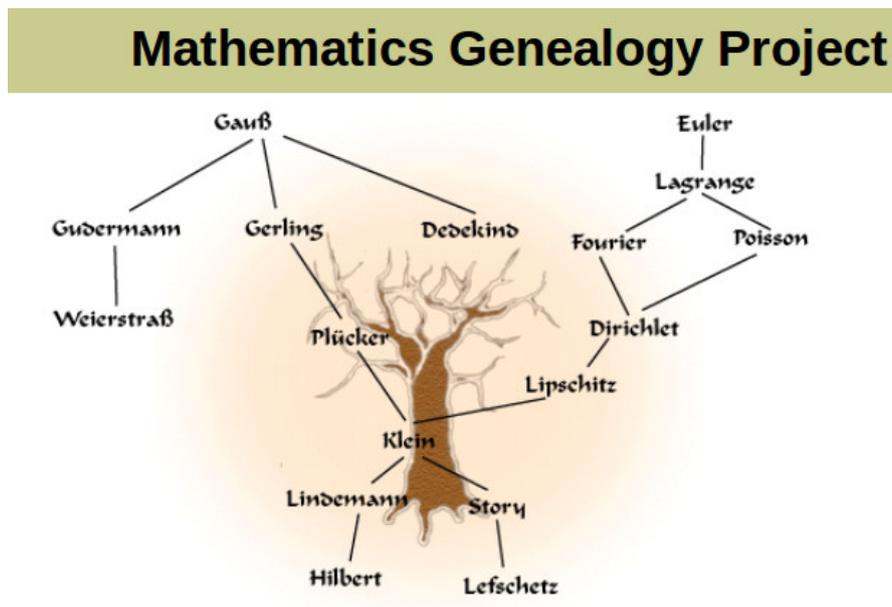
6.4 Future Work

This work shows that the quality of the results of a Data Science-based analysis of Lattes CVs relies on the quality of the data presented in these CVs. For that, some actions may be taken to improve the reliability of the data:

- **An implementation of disambiguation techniques:** it was observed in this work that, throughout the Lattes CVs published data, one researcher may be referred to by different nomenclatures, depending on what parts of their name are abbreviated or simply not cited by the papers. To avoid or reduce the quantity of errors caused by this condition, a disambiguation algorithm should be used for the cases where analysis retrieves co-authors information from the Lattes CV. (RUBIM; BRAGANHOLO, 2017) has conducted a thorough analysis on disambiguation errors on Lattes CVs, that can be used to verify the same results in LattesLab.
- **Data comparison / Information update:** in the previous topic, the subject of disambiguation of author's name was raised. But that is not the only information in LattesLab subject to error or to be out of date. In the analysis of the Lattes CVs, a relationship of the researchers - based on their history of publishing together - has been found. So, if two researchers publish a paper together and only one of them updated their Lattes CV, it is possible to identify that the other Lattes CV is out-of-date, and even suggest possible editions to the CV.

- **Generate Advisers Genealogy trees:** Through Academic Genealogy trees, it is possible to identify relationships between researchers and advisers and the advisers of their advisers, to establish historical connection. One of the most known examples is a mathematics genealogy tree that contains several renowned mathematicians, as found in figure 6.1. Through LattesLab, it is also possible to generate social networks based on researchers that have oriented other researchers. There's a whole body of research in this subject, some of them using the Lattes CV as a data source (WANG et al., 2010) (DORES et al., 2017) (MOREIRA et al., 2017). In order to verify if LattesLab is a valid tool for this application, the current solutions can be analyzed and compared with LattesLab functionalities. It is also necessary to validate if these functionalities need to be extended for a valid, scientific analysis of these academic genealogy trees.

Figure 6.1 - Mathematics Genealogy Tree.

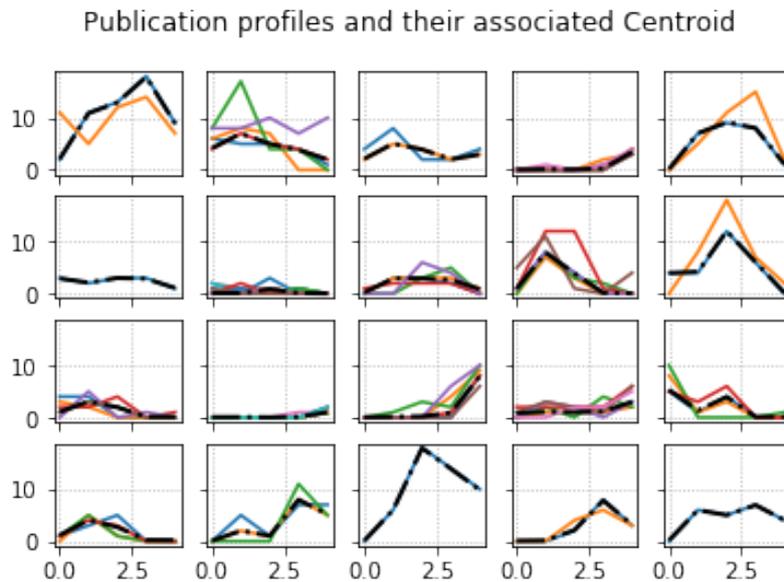


SOURCE: Mathematics Genealogy Project (2018).

- **A new proposed model of researcher's publications:** it has been seen that the assumption of a twenty-year interval of publication per researcher influences the unsupervised training models in a way that harms

the minority of researchers that have in fact published for that period – since most of the researchers don't. The figure 6.2 shows how different the results from section 5.3.1 are when a vector of five years of publications – instead of twenty years – is used.

Figure 6.2 - Fuzzy C-Means model with 20 centroids and publications spanning through five years.



SOURCE: LattesLab library output.

Therefore, a way to remove this sensibility to the span of years of researchers should be found – or at least sought. Some possibilities are:

- Use a variable containing the number of years that the researcher has published.
- Use the mean of publications per year and a standard deviation.
- Verify if the researcher has pursued post-graduation studies and use this information to refine the analysis.

REFERENCES

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. Lattesminer: a multilingual dsl for information extraction from lattes platform. In: ACM WORKSHOP ON DSM, 2011. **Proceedings...** [S.l.], 2011. p. 85–92. 18

_____. Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 2011. **Proceedings...** [S.l.], 2011. p. 1–6. 9, 10, 18

ALVES, W. A.; SANTOS, S. D.; SCHIMIT, P. H. Hierarchical clustering based on reports generated by scriptlattes. In: IFIP INTERNATIONAL CONFERENCE ON ADVANCES IN PRODUCTION MANAGEMENT SYSTEMS, 2016. **Proceedings...** [S.l.], 2016. p. 28–35. 17

ARAÚJO, E. B.; MOREIRA, A. A.; FURTADO, V.; PEQUENO, T. H.; JR, J. S. A. Collaboration networks from a large cv database: dynamics, topology and bonus impact. **PloS one**, v. 9, n. 3, p. e90537, 2014. 9, 10, 19

BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: the fuzzy c-means clustering algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191–203, 1984. 49

BONIFACIO, A. S. **Ontologias e consulta semântica: uma aplicação ao caso lattes. 2002.** 85 p. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. 19

BRASIL. SENADO FEDERAL. Investimento em pesquisa e desenvolvimento no brasil e em outros países: o setor privado. **Em Discussão!**, v. 3, n. 12, p. 25, 2012. Disponível em: <<https://bit.ly/2NC8Z8F>>. 1, 4

BUREAU OF LABOR STATISTICS. **Unemployment rates and earnings by educational attainment.** 2018. Disponível em: <https://www.bls.gov/emp/ep_chart_001.htm>. 3

CENTRA, J. A. Research productivity and teaching effectiveness. **ETS Research Report Series**, v. 1981, n. 1, p. i–14, 1981. ISSN 2330-8516. Disponível em: <<http://dx.doi.org/10.1002/j.2333-8504.1981.tb01246.x>>. 6

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ. **Extração de dados - Plataforma Lattes - CNPq.**

2017. Disponível em:

<<http://lattes.cnpq.br/web/plataforma-lattes/extracao-de-dados/>>. 10

_____. **Bolsas no País: Investimentos realizados segundo modalidades.**

2018. Disponível em: <<http://cnpq.br/documents/10157/cd16cb40-a092-4d19-a3fb-f6e85e8c8f80>>. 30

_____. **Lattes CV XSD file.** 2018. Disponível em: <<http://memoria.cnpq.br/documents/313759/dce5f2c1-0675-4907-be58-4442c21c6333>>. 14

_____. **Lattes CV XSD file.** 2018. Disponível em: <<http://memoria.cnpq.br/documents/313759/dce5f2c1-0675-4907-be58-4442c21c6333>>. 14

_____. **Painel de investimentos.** 2018. Disponível em:

<<http://memoria.cnpq.br/painel-de-investimentos>>. 2

_____. **Séries históricas de investimento em pesquisa.** 2018. Disponível em:

<<http://cnpq.br/documents/10157/87141ebd-81d0-4243-85f8-e817737b61b1>>. 2

_____. **Séries históricas de investimento em pesquisa.** 2018. Disponível em:

_____. **Total of investments in scholarships and promotion of research - 1996-2015.** 2018. Disponível em: <<http://www.cnpq.br/documents/10157/87141ebd-81d0-4243-85f8-e817737b61b1>>. 2

_____. **Total of investments in scholarships and promotion of research - 1996-2015.** 2018. Disponível em: <<http://www.cnpq.br/documents/10157/87141ebd-81d0-4243-85f8-e817737b61b1>>. 2

DIGIAMPIETRI, L.; MENA-CHALCO, J.; PÉREZ-ALCÁZAR, J. J.; TUESTA, E. F.; DELGADO, K.; MUGNAINI, R. Minerando e caracterizando dados de currículos lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2012. **Proceedings...** [S.l.], 2012. 19

DORES, W.; SOARES, E.; BENEVENUTO, F.; LAENDER, A. H. Building the brazilian academic genealogy tree. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 2017. **Proceedings...** [S.l.], 2017. p. 537–543. 68

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. In: . [s.n.], 1973. v. 3, n. 3, p. 32–57. Disponível em: <<https://doi.org/10.1080/01969727308546046>>. 49

FERNANDES, G. O.; SAMPAIO, J. O.; SOUZA, J. Xmlattes a tool for importing and exporting curricula data. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE ENGINEERING, 2011. **Proceedings...** [S.l.], 2011. 18

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO –
FAPESP. **Scientific Electronic Library Online**. 2018. Disponível em:
<<http://www.scielo.org/php/index.php>>. 6

GALEGO, E. F. **Extração e consulta de informações do Currículo Lattes baseada em ontologias**. 2013. 70 p. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, São Paulo, 2013. 19

GOOGLE. **Google Scholar**. 2018. Disponível em:
<<https://scholar.google.com>>. 6

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: a k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100–108, 1979. 49

HEIMERL, F.; LOHMANN, S.; LANGE, S.; ERTL, T. Word cloud explorer: text analytics based on word clouds. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 47., 2014. **Proceedings...** [S.l.]: IEEE, 2014. p. 1833–1842. 41

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, v. 102, n. 46, p. 16569–16572, 2005. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283832/pdf/pnas-0507655102.pdf>>. 6

IHAKA, R.; GENTLEMAN, R. R. a language for data analysis and graphics. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 299–314, 1996. 23

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys (CSUR)**, v. 31, n. 3, p. 264–323, 1999. 47

JOHNSSON, M. **Applications of self-organizing maps**. [S.l.]: InTech, 2012. 50

KAY, A. Tesseract: an open-source optical character recognition engine. **Linux Journal**, v. 2007, n. 159, p. 2, 2007. 18

KLUYVER, T. et al. Jupyter notebooks—a publishing format for reproducible computational workflows. **Positioning and Power in Academic Publishing: Players, Agents and Agendas**, p. 87, 2016. 24

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, n. 1, p. 59–69, 1982. 50

LIU, X.; BOLLEN, J.; NELSON, M. L.; SOMPEL, H. Van de. Co-authorship networks in the digital library research community. **Information processing & management**, v. 41, n. 6, p. 1462–1480, 2005. 44

MARQUES, K.; ODDONE, N.; MIRANDA, M. Organização da informação na plataforma lattes. 2007. Disponível em:
<<http://www.enancib.ppgci.ufba.br/artigos/GT2--338.pdf>>. 11

MATHEMATICS GENEALOGY PROJECT. 2018. Disponível em:
<<https://www.genealogy.math.ndsu.nodak.edu/index.php>>. 68

MENA-CHALCO, J. P.; CESAR JUNIOR, R. M. Scriptlattes: an open-source knowledge extraction system from the lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, Dec 2009. ISSN 1678-4804. Disponível em: <<https://doi.org/10.1007/BF03194511>>. 10, 17

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; LOPES, F. M.; CESAR, R. M. Brazilian bibliometric coauthorship networks. **Journal of the Association for Information Science and Technology**, v. 65, n. 7, p. 1424–1445, 2014. 9, 10, 19

MITCHELL, J. C. **Social networks in urban situations**. [S.l.]: Manchester University Press for the Institute for African Studies, University of Zambia, 1969. 31

MOREIRA, T. H.; DIAS, T. M.; MOITA, G. F. Genealogia acadêmica da relação orientador-orientado na área de ciência da computação. **Anais do Computer on the Beach**, p. 060–069, 2017. 68

NAKASHIMA, M. Y. **Curriculo Lattes e Web Semantica**. 2004. Disponível em:
<<https://www.linux.ime.usp.br/~cef/mac499-04/monografias/rec/myn/>>. 19

NOY, N. F. et al. **Ontology development 101: A guide to creating your first ontology**. [S.l.]: Stanford: Sanford University, 2001. 19

PENG, R. D. Reproducible research in computational science. In: . [S.l.: s.n.], 2011. v. 334, n. 6060, p. 1226–1227. 22, 23

PEREZ-CERVANTES, E.; MENA-CHALCO, J. P.; CESAR, R. M. Towards a quantitative academic internationalization assessment of brazilian research groups. In: INTERNATIONAL CONFERENCE ON E-SCIENCE, 8., 2012. **Proceedings...** [S.l.]: IEEE, 2012. p. 1–8. 9, 17, 19

RANKING WEB OF UNIVERSITIES. **2258 highly cited researchers (h>100) according to their Google scholar citations public profiles**. 2018. Disponível em: <<http://www.webometrics.info/en/node/58>>. 6

RUBIM, I. C.; BRAGANHOLO, V. Detecting referential inconsistencies in electronic cv datasets. **Journal of the Brazilian Computer Society**, v. 23, n. 1, p. 3, 2017. 67

SANTANA, T. L. V.; SANTOS, R. Data science approach to analysis of lattes cv data. **Central Europe Ceur Workshop Proceedings**, v. 2029, p. 168–177, 2017. 67

SCHUTT, R.; O'NEIL, C. **Doing data science: straight talk from the frontline**. [S.l.]: O'Reilly Media, 2013. 21, 22

SEMATECH e-handbook of statistical methods. 2013. Disponível em: <<http://www.itl.nist.gov/div898/handbook/>>. 31

TORRESI, S. I. C. A.; PARDINI, V. L.; FERREIRA, V. F. Fraudes, plágios e currículos. **Química Nova**, v. 32, p. 1371 – 1371, 00 2009. ISSN 0100-4042. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422009000600001&nrm=iso>. 11

UNITED NATIONS - UN. **International system for agricultural science and technology**. 2018. Disponível em: <<http://agris.fao.org/agris-search/index.do>>. 6

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - UFRN. **Cursos avaliados e reconhecidos - INPE**. 2018. Disponível em: <<https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/quantitativos/quantitativoPrograma.jsf?areaAvaliacao=0&cdRegiao=3&sgUf=SP&ies=338723>>. 4

UNIVERSITY WORLD NEWS. **The Role of Research Universities in Developing Countries - University World News₂₀₁₈**. 2018. Disponível em: <>. 2

VANROSSUM, G.; DRAKE, F. L. **The python language reference**. [S.l.]: Amsterdam: Python Software Foundation, 2010. 23

W3SCHOOLS. **Introduction to XML**. 2018. Disponível em: <https://www.w3schools.com/xml/xml_what_is.asp>. 12

_____. **XML Attributes**. 2018. Disponível em: <https://www.w3schools.com/xml/xml_attributes.asp>. 13

WANG, C.; HAN, J.; JIA, Y.; TANG, J.; ZHANG, D.; YU, Y.; GUO, J. Mining advisor-advisee relationships from research publication networks. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 16, 2010. **Proceedings...** [S.l.]: ACM, 2010. p. 203–212. 68

WORLD BANK. **Brazil's GDP**. 2018. Disponível em: <<https://data.worldbank.org/country/brazil>>. 3

YAN, S.; XU, D.; ZHANG, B.; ZHANG, H.-J.; YANG, Q.; LIN, S. Graph embedding and extensions: A general framework for dimensionality reduction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 29, n. 1, p. 40–51, 2007. 47