astrogen

Release 0.1

Marcelo Lares

Dec 08, 2021

PROJECT

ASTROGEN is a project devoted to the formal statistical analysis of the gender representations in the astronomical community in Argentina.

CHAPTER

ONE

PROPOSAL

In this project we propose to carry out a datat intensive analysis of the situation in the Argentina astronomical workforce regarding the gender balance.

We make use of several tools and techniques, including:

- hypothesis testing
- machine learning
- word embeddings
- web scraping
- statistical analysis
- model selection and assessment

We aim at leveraging open data to produce our results, and we publish the full stack of the data reduction and analysis pipeline so that anyone can use the data and reproduce the results or revamp the analysis.

This way, our results and conclusions can be revised by researchers in the data science or gendeer studies communities.

1.1 Methods

We propose the following working hypothesis: In Argentina, in the profesional astronomical community, there is a segregation effect in gender due to structural causes and independent of personal choices. This segregation or gender bias leads to a noticeable unfair availability of opportunities in the development of careers for women and men.

1.2 Data analysis

Statistical analysis of the following aspects on the astronomy career in Argentina:

- Gender representation in the academic career in Universities
- Gender representation in scholarships (CONICET)
- Gender representation in permanent positions (CIC, CONICET)
- Gender differences in scientific production metrics
- Gender differences in the academic performance in Universities

CHAPTER

TWO

DATA PIPELINE

(these links are temporarily restricted to authors)

- Shared Drive with data warehouse (requires access rights)
- Github repository for codes
- Overleaf document for paper (requires access rights)
- Data repository (comming soon)

2.1 Get the data

The data used in this work has been collected from several sources, namely:

- Astronomical Data System
- CONICET
- CONICET, "gobierno abierto > conicet en cifras" https://cifras.conicet.gov.ar/publica/
- CONICET, "conicet digital" Repositorio institucional https://ri.conicet.gov.ar
- Asociación Argentina de Astronomía
- Astronomy Institutes in Argentina:
  - IATE
  - IAFE
  - ICATE
  - IALP
  - OAC
- Universities in Argentina
  - **UNC**
    - Lic. en Astronomía
    - Statistical data

2.2 Feature construction

- Age: We performed a non-linear least squares regression for the variables "DNI" and age, using as the training dataset that of the AAA original table.
- Gender: We use the table compiled by Mustafa Atik to assign gender on the basis of the names. We have also tried other tools, e.g. genderize.io, throught the client https://github.com/SteelPangolin/genderize, GenderAPI web tool, with the same results.

2.3 Publications

We have performed a detailed analysis of publication data for each author. The process involves the following steps:

- Download ADS publication data for each author using the name as the search key. At this stage we use the python packages ADS and PINNACLE.
- Search and add ORCID keys
- Train, evaluate and apply a machine learning model to clean the sample of papers. This is required since the search by name often return the entries for several authors with similar names.
- Add metrics for journals, taking data from SCImago Journal & Country Rank public portal
- Add publication metrics

The curated dataset allows to construct several indices:

- number of authors per article
- number of articles per author
- number of articles per author, as leading author
- distribution in time of articles
- H-index
- relative position of a given author in the authors list

CHAPTER

THREE

RESEARCH OUTPUTS

3.1 Meetings

- Congreso de Ciencia y Género Eje 7D: Paridad y participación en ciencia y ámbitos científicos. Presentación: "Análisis estadístico de la paridad de género en la Astronomía Argentina", M. Lares, V. Coenda, L. Gramajo, J. Martínez Atencio, C. Parisi, C. Ragone Figueroa, oct. 1, 2021.

3.2 Articles

- Is professional astronomy in Argentina an unicorn ingender equality? (in preparation)

3.3 Media

CHAPTER

FOUR

DATASET

GENDER BALANCE IN THE ARGENTINA ASTRONOMY WORKFORCE

This dataset is published in http://dryad/datasets/astrogen, see that link for full access to the data.

4.1 Metadata

Dataset compiled from several oficial and public sources about the career development for astronomers in Argentina.

- Marcelo Lares [1, 2, 3] ORCID:
- Valeria Coenda [1, 2, 3] ORCID:
- Luciana Gramajo [1, 2, 3] ORCID:
- Héctor Julián Martínez-Atencio [1, 2, 3] ORCID:
- Celeste Parisi [1, 2, 3] ORCID:
- Cinthia Ragone [1, 3] ORCID:

Affiliations:

- 1) Instituto de Astronomía Teórica y Experimental (IATE)
- 2) Observatorio Astronómico de Córdoba (OAC)
- 3) CONICET

Contact: Marcelo Lares

Date of data collection: Nov 29, 2021

GEOGRAPHIC LOCATION: Argentina

KEYWORDS: gender balance, astronomy

LANGUAGE: English

Funding sources: The authors acknowledge founding from CONICET and SECYT, although to granted for this project especifically.

4.2 Data and file overview

We provide a single file containing an SQL database with five tables:

table	#elements	#columns
famaf	210	6
people	838	19
papers	341825	9

- famaf

Count of male and female students, by year and year of enrollment.

Column	Name	format	description
1	year	INT	year the count of students is made
2	year_in	INT	year of enrollment
3	mi	INT	number of male active students in year "year" that enrolled in year "year_in".
4	me	INT	number of male students that obtain the degree in year "year" and enrolled in year "year_in".
5	fi	INT	number of female active students in year "year" that enrolled in year "year_in".
6	fe	INT	number of female students that obtain the degree in year "year" and enrolled in year "year_in".

- people

List of astronomers in Argentina

Column	Name	format	contents
1	Author ID	INT	Unique identifier
2	age	INT	age [years]
3	gender	CHAR	gender (m, f)
4	Hindex	INT	H-index for publication set
5	Npapers	INT	number of papers
6	cc07	INT	category in CONICET in 2007
7	cc08	INT	category in CONICET in 2008
8	cc09	INT	category in CONICET in 2009
9	cc10	INT	category in CONICET in 2010
10	cc11	INT	category in CONICET in 2011
11	cc12	INT	category in CONICET in 2012
12	cc13	INT	category in CONICET in 2013
13	cc14	INT	category in CONICET in 2014
14	cc15	INT	category in CONICET in 2015
15	cc16	INT	category in CONICET in 2016
16	cc17	INT	category in CONICET in 2017
17	cc18	INT	category in CONICET in 2018
18	cc19	INT	category in CONICET in 2019
19	cc20	INT	category in CONICET in 2020

- papers

List of papers

Column	Name	format	contents
1	ID	INT	Author identifier
2	journal	INT	journal name
3	journal_Q	INT	Q index for journal (from SCIMAGO). 0: not indexed, 1: first quartile, 2: second quartile, 3: third quartile, 4: fourth quartile
4	year	INT	year of the publication
5	citation_count	INT	number of citations (at the date of compilation)
6	author_count	INT	number of authors
7	author_pos	INT	position of author in author list
8	inar	INT	identifier of author affiliation. 0: not in Argentina, 1: in Argentina, 2: not declared.
9	filter	INT	automatic filter. 0: do not belong to the author, 1: assigned to the author

The papers have been classified by an automatic agent as belonging to the author. The full set of publications retrieved from the ADS service is classified according to this classifier, which gives the "filter" column as a result.

The fields ID allows to relate the tables "people" and "papers".

4.3 Sample selections

We use a subset from the "people" table, corresponding to authors tha satisfy the following criteria:

- Active on 2021 (last published paper in a Q1 journal and not from a large collaboration not before 2016)
- Age in the range 25 to 85 years old
- At least 75% of the Q1 papers (excluding large collaborations) published with an affiliation in Argentina

This dataset can be obtained from the database using, for example, the following SQL query:

```
select *,
       COUNT(*) as cc,
       MAX(p.year) as ymx,
       SUM(CASE WHEN p.inar=1 then 1 else 0 END) as N_inar,
       SUM(CASE WHEN p.inar=1 then 1 else 0 END) / (1.*COUNT(*)) as q
FROM papers as p
    INNER JOIN people as g
WHERE
        p.ID==g.ID
        AND
        g.age BETWEEN 25 AND 85
        AND
        p.journal_Q==1
        AND
        p.author_count<51
GROUP BY p.ID
HAVING
   ymx>2016
   AND
   q>0.75
```

Another subset is the one that comprise all researchers in CONICET at a given year.

The following query returns a subset from the "people" table corrresponding to active researchers at CONICET in 2020:

```
select * from people
where cc20 is not NULL
```

The list of publications from a given author can be obtained using the ID fields. For example, all the publications in top journals from the author with ID=35 can be obtained as:

```
select * from papers
where
  ID==35
  and
  journal_Q==1
```

SQL queries can be easily run, either using apropriate software (e.g. DB browser for sqlite) or using sqlite3 in python.

Let be the *astrogen* object a string containing the root directory of the project. The following code allows to read all researchers that where category I in 2015 and category II in 2020.

```
from os import path
from sqlite3 import connect

db = path.join(astrogen, 'data/redux/astrogen_DB_anonymous.db')
conn = connect(db)
c = conn.cursor()
query = ('''
          select *
          from people
          where
            cc17==1
            AND
            cc20==2
          ''')
c.execute(query)
df = pd.DataFrame(c.fetchall())
conn.close()
```

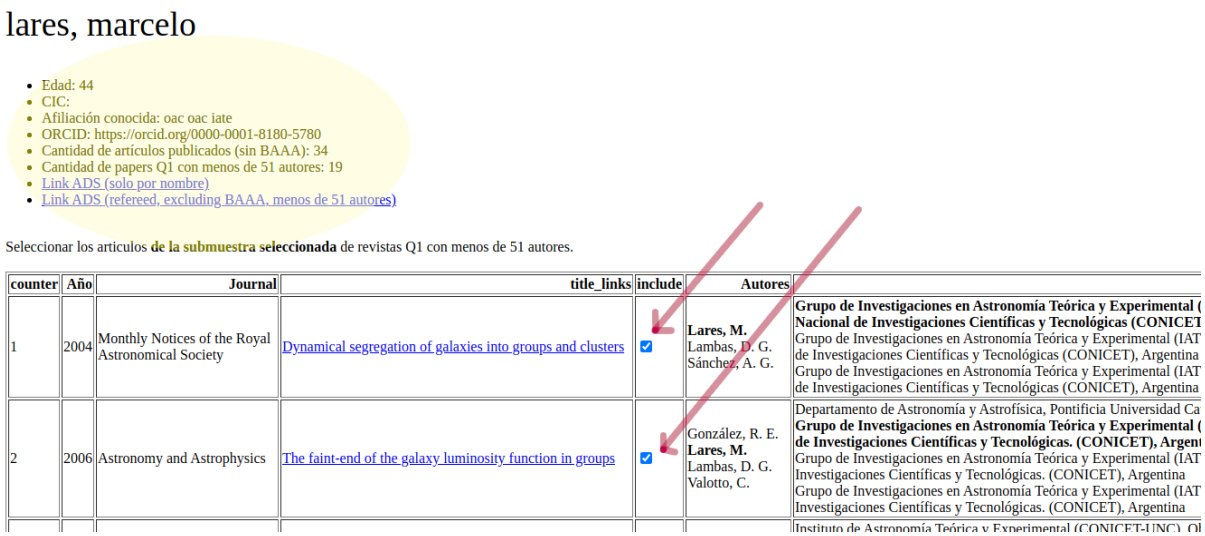
4.4 Validation of the publication lists

Once the list of publications retrieved from the Astronomical Data Service (ADS) has been classified using a Support Vector Machine model, we prepare pages for each author in order to visually verify possible sources or error.

A sample page can be found here. In these pages we include a link to the ADS entry on the author, using the same search string that was used when retrieving information from the ADS server using the ADS python package.

We also include a link to the ADS page of the author requesting only papers with at most 50 authors and excluding the BAAA publication, which is a proceeding from the AAA annual meetings that authors use to have many entries.

The check marks are the automatic selection made by the classifier, and these pages allow to correct false positives or false negatives by changing the tickmarks and saving a file with the updated filter.



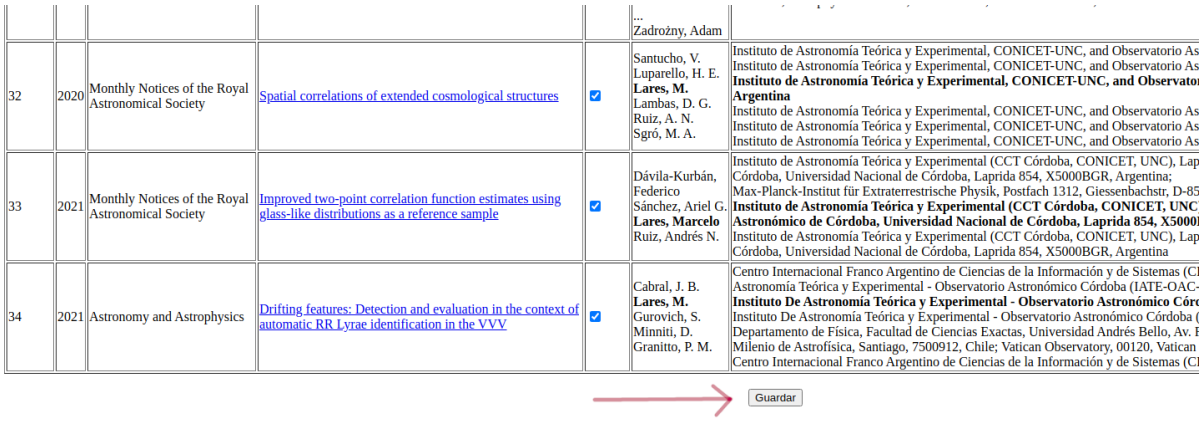
lares, marcelo

- Edad: 44
- CIC:
- Afiliación conocida: oac oac iate
- ORCID: https://orcid.org/0000-0001-8180-5780
- Cantidad de artículos publicados (sin BAAA): 34
- Cantidad de papers Q1 con menos de 51 autores: 19
- Link ADS (solo por nombre)
- Link ADS (refereed, excluding BAAA, menos de 51 autores)

Seleccionar los articulos **de la submuestra seleccionada** de revistas Q1 con menos de 51 autores.

counter	Año	Journal	title_links	include	Autores	
1	2004	Monthly Notices of the Royal Astronomical Society	Dynamical segregation of galaxies into groups and clusters	☑	**Lares, M.** Lambas, D. G. Sánchez, A. G.	**Grupo de Investigaciones en Astronomía Teórica y Experimental (Nacional de Investigaciones Científicas y Tecnológicas (CONICET** Grupo de Investigaciones en Astronomía Teórica y Experimental (IAT de Investigaciones Científicas y Tecnológicas (CONICET), Argentina Grupo de Investigaciones en Astronomía Teórica y Experimental (IAT de Investigaciones Científicas y Tecnológicas (CONICET), Argentina
2	2006	Astronomy and Astrophysics	The faint-end of the galaxy luminosity function in groups	☑	González, R. E. **Lares, M.** Lambas, D. G. Valotto, C.	Departamento de Astronomía y Astrofísica, Pontificia Universidad Ca **Grupo de Investigaciones en Astronomía Teórica y Experimental (de Investigaciones Científicas y Tecnológicas. (CONICET), Argen** Grupo de Investigaciones en Astronomía Teórica y Experimental (IAT Investigaciones Científicas y Tecnológicas. (CONICET), Argentina Grupo de Investigaciones en Astronomía Teórica y Experimental (IAT Investigaciones Científicas y Tecnológicas. (CONICET), Argentina
						Instituto de Astronomía Teórica y Experimental (CONICET-UNC), Ol

Example of the saving button:



					... Zadrożny, Adam	
32	2020	Monthly Notices of the Royal Astronomical Society	Spatial correlations of extended cosmological structures	☑	Santucho, V. Luparello, H. E. **Lares, M.** Lambas, D. G. Ruiz, A. N. Sgró, M. A.	Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observatorio As Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observatorio As **Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observato Argentina** Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observatorio As Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observatorio As Instituto de Astronomía Teórica y Experimental, CONICET-UNC, and Observatorio As
33	2021	Monthly Notices of the Royal Astronomical Society	Improved two-point correlation function estimates using glass-like distributions as a reference sample	☑	Dávila-Kurbán, Federico Sánchez, Ariel G. **Lares, Marcelo** Ruiz, Andrés N.	Instituto de Astronomía Teórica y Experimental (CCT Córdoba, CONICET, UNC), Lap Córdoba, Universidad Nacional de Córdoba, Laprida 854, X5000BGR, Argentina; Max-Planck-Institut für Extraterrestrische Physik, Postfach 1312, Giessenbachstr, D-85 **Instituto de Astronomía Teórica y Experimental (CCT Córdoba, CONICET, UNC Astronómico de Córdoba, Universidad Nacional de Córdoba, Laprida 854, X5000** Instituto de Astronomía Teórica y Experimental (CCT Córdoba, CONICET, UNC), Lap Córdoba, Universidad Nacional de Córdoba, Laprida 854, X5000BGR, Argentina
34	2021	Astronomy and Astrophysics	Drifting features: Detection and evaluation in the context of automatic RR Lyrae identification in the VVV	☑	Cabral, J. B. **Lares, M.** Gurovich, S. Minniti, D. Granitto, P. M.	Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CI Astronomía Teórica y Experimental - Observatorio Astronómico Córdoba (IATE-OAC- **Instituto De Astronomía Teórica y Experimental - Observatorio Astronómico Córd** Instituto De Astronomía Teórica y Experimental - Observatorio Astronómico Córdoba (Departamento de Física, Facultad de Ciencias Exactas, Universidad Andrés Bello, Av. F Milenio de Astrofísica, Santiago, 7500912, Chile; Vatican Observatory, 00120, Vatican Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CI

Guardar

4.5 References

FILE FORMATS. Cornell Research Data Management Service Group. http://data.research.cornell.edu/content/file-formats

FILE MANAGEMENT. Cornell Research Data Management Service Group. http://data.research.cornell.edu/content/file-management

CHAPTER

FIVE

ABOUT US

We are a group of astronomers interested in highlighting the value of complete, curated and unbiased data to discuss gender topics in the astronomical community.

Authors:

- Valeria Coenda [1,2]
- Luciana Gramajo [1,2]
- Marcelo Lares [1,2]
- Julián Martínez [1,2]
- Celeste Parisi [1,2]
- Cinthia Ragone [1]

Affilliations:

1. Instituto de Astronomía Teórica y Experimental (CONICET - UNC)
2. Observatorio Astronómico de Córdoba (UNC)

Contact: marcelo.lares@unc.edu.ar

Here we describe the tools (software) to retrieve, curate and analyze data related to the project.

CHAPTER

SIX

ASTROGEN_UTILS MODULE

This module contains useful functions to manipulate strings for name matching, pretty printing, gender detection and XLSX output.

The functions in this module comprise:

- **string comparisson:**
  - ds
  - ds1
  - ds2
  - initials
  - getinitials
  - pickone
- **input/output:**
  - bcolors
  - append_df_to_excel

astrogen_utils.**append_df_to_excel**(*filename*, *df*, *sheet_name='Sheet1'*, *startrow=None*, *truncate_sheet=False*, ***to_excel_kwargs*)

Append a DataFrame [df] to existing Excel file [filename] into [sheet_name] Sheet. If [filename] doesn't exist, then this function will create it.

@param filename: File path or existing ExcelWriter (Example: '/path/to/file.xlsx') @param df: DataFrame to save to workbook @param sheet_name: Name of sheet which will contain DataFrame. (default: 'Sheet1') @param startrow: upper left cell row to dump data frame. Per default (startrow=None) calculate the last row in the existing DF and write to the next row... @param truncate_sheet: truncate (remove and recreate) [sheet_name] before writing DataFrame to Excel file @param to_excel_kwargs: arguments which will be passed to *DataFrame.to_excel()* [can be a dictionary] @return: None

Usage examples:

```
>>> append_df_to_excel('d:/temp/test.xlsx', df)
```

```
>>> append_df_to_excel('d:/temp/test.xlsx', df, header=None, index=False)
```

```
>>> append_df_to_excel('d:/temp/test.xlsx', df, sheet_name='Sheet2',
                       index=False)
```

```
>>> append_df_to_excel('d:/temp/test.xlsx', df, sheet_name='Sheet2',
                       index=False, startrow=25)
```

(c) [MaxU](https://stackoverflow.com/users/5741205/maxu?tab=profile)

class astrogen_utils.**bcolors**

Bases: object

Get color palette for pretty printing

This class simply contains a list of predefined colors to be used in the visual analysis of strings and publication data.

BOLD = '\x1b[1m'

ENDC = '\x1b[0m'

FAIL = '\x1b[91m'

HEADER = '\x1b[95m'

OKBLUE = '\x1b[94m'

OKCYAN = '\x1b[96m'

OKGREEN = '\x1b[92m'

TST = '\x1b[31;1m'

UNDERLINE = '\x1b[4m'

WARNING = '\x1b[93m'

X = '\x1b[4;95;1m'

astrogen_utils.**clean_text**(*txt*)

astrogen_utils.**df_to_dict**(*df*, *key_column*, *val_column*)

convierte dos pandas series en un diccionario

astrogen_utils.**ds**(*a*, *b*)

Get distance between two words.

This function is used to obtain the distance between two names or surnames. Uses different distances in word space, namely, Damerau Levenshtein distance, Jaro distance, Levenstein distance and SequenceMatcher. The later from the difflib package and the other ones from the Jellyfish package.

Args: a (string): one of the strings b (string): the other string to compare

Returns:

res (array): Numpy array with the list of distances between the two words.

astrogen_utils.**ds1**(*s1*, *s2*)

Get distance between two words.

This function is used to obtain the distance between two names or surnames. Uses different distances in word space, namely, Damerau Levenshtein distance, Jaro distance, Levenstein distance and SequenceMatcher. The later from the difflib package and the other ones from the Jellyfish package.

Args: a (string): one of the strings b (string): the other string to compare

Returns:

res (array): Numpy array with the list of distances between the two words.

astrogen_utils.**ds2**(*ap1*, *ap2*, *nom1*, *nom2*)
Get distance between two words.

This function is used to obtain the distance between two names or surnames. Uses different distances in word space, namely, Damerau Levenshtein distance, Jaro distance, Levenstein distance and SequenceMatcher. The later from the difflib package and the other ones from the Jellyfish package.

Args: a (string): one of the strings b (string): the other string to compare

Returns:

res (array): Numpy array with the list of distances between the two words.

astrogen_utils.**fnames**(*auth*, *folder*, *extension*, *include_path=True*)
build the file name

astrogen_utils.**get_gender2**(*names*)

astrogen_utils.**getinitials**(*nombre*)
Get the initials of a full name

e.g.: 'Jose Facundo' –> 'J. F.'

Args:

Returns:

astrogen_utils.**getinitialscompact**(*nombre*)
Get the initials of a full name

e.g.: 'Jose Facundo' –> 'JF'

Args:

Returns:

astrogen_utils.**initials**(*initials*, *string*)
Check if the initials of two names coincide.

e.g.:

initials = 'Juan Carlos'; string='Juan' –> True

initials = 'Juan Carlos'; string='Juan José' –> False

initials = 'Juan Carlos'; string='Jacinto' –> True

Args: initials (string): source string for the initials string (string): full names

Returns: boo (bool): whether the initials are accepted

Notes:

The criteria for the string matching is the following:

astrogen_utils.**pickone**(*df*, *au*, *sift*)
de una lista de autores en un dataframe "df" elige el que está más cerca de un autor "au" y devuelve un array booleano que es todo falso savo uno (el autor elegido).

astrogen_utils.**similar**(*a*, *b*)

CHAPTER

SEVEN

PIPELINE MODULE

The data reduction pipeline is implemented through a bonobo pipeline, within an ETL (extract-transform-load) model, with the following steps:

action	routine
read table from AAA	• `pipeline.S01_read_aaa_table()`
merge with tables from institutes	• `pipeline.S02_add_OAC_data()` • `pipeline.S02_add_IATE_data()` • `pipeline.S02_add_IALP_data()` • `pipeline.S02_add_ICATE_data()` • `pipeline.S02_add_GAE_data()`
merge with data from CONICET	• `pipeline.S02_add_CONICET_data()`
add gender	• `pipeline.S03_add_gender()`
add age	• `pipeline.S03_add_age()`
clean papers	• `pipeline.S04_pub_get_ads_entries()` • `pipeline.S04_pub_get_orcids()` • `pipeline.S04_pub_journal_index()`
add journal index	• `pipeline.S04_pub_clean_papers()`
add publication metrics	• `pipeline.S04_pub_value_added()`
visual check	• `pipeline.S04_make_pages()` • `pipeline.S04_load_check_filters()`
anonymize	• `pipeline.S05_anonymize()`

In what follows we describe each step separately.

The module `pipeline()` contains the steps fot the data reduction pipeline.

The steps are

- S01: read base table (AAA)
- **S02: add institutes and cic data** In these steps the following columns are added:
  - cic
  - docencia
  - area
  - orcid
  - use_orcid

  The steps are contained in the following functions:
  - `pipeline.S02_add_OAC_data()`
  - `pipeline.S02_add_IATE_data()`
  - `pipeline.S02_add_IALP_data()`
  - `pipeline.S02_add_ICATE_data()`
  - `pipeline.S02_add_GAE_data()`
  - `pipeline.S02_add_CIC_data()`
- **S03: add metadata for authors**
  - S03_add_gender
  - S03_add_age
  - S03_clean_and_sor
- **S04: add publications data**
  - `pipeline.S04_pub_get_ads_entries()`
  - `pipeline.S04_pub_get_orcids()`
  - `pipeline.S04_pub_journal_index()`
  - `pipeline.S04_pub_clean_papers()`
  - `pipeline.S04_make_pages()`
  - `pipeline.S04_pub_value_added()`

API documentation for the code in **pipeline.py**:

CHAPTER

EIGHT

EXAMPLE

This project is organized as an API to be used from a python prompt.

Steps:

- Complete the configuration of the experiment
- All the settings of the experimets are parsed from the configuration files using configparser.

8.1 Installation

First, download the latest version of the code repository ASTROGEN in GitHub.

The code has been tested in the following python versions:

- 3.8.5

The list of requirements is given in the file "requirements.txt":

- numpy
- scipy
- ads
- sklearn

In order to run the data pipeline, which connects to the ADS online database, an API KEY is required. Documentation for obtaining and using this key can be found in the ADS API documentation.

```
$ conda create --name astrogen ads==0.12.3 bonobo==0.6.4 docutils==0.17.1 jellyfish==0.8.
↪8 joblib==1.1.0 matplotlib==3.4.3 numpy==1.21.2 openpyxl==3.0.9 pandas==1.3.3 scikit-
↪learn==1.0 scipy==1.7.1
```

However, the reccommended method is to create a virtual environment (using either conda or virtualenv) and then:

```
pip install -r requirements.txt
```

8.2 Configuration

The main configurations are set in a configuration file writen in a configuration file.

An example of the configuration file is as follows:

```
# ____________________________________________________
[experiment] # EXPERIMENT ID

# Experiment ID.  Useful to compare and save experiments.
# A directory will be created with this name under [out]dir_output
experiment_ID = TRNT_001

# ____________________________________________________
[dirs] # Directory structure (relative to: astrogen/)

# locations of data files
datadir_root = data/

# locations of external data files
# relative to $datadir_root
datadir_external = external

# locations of interim data files
# relative to $datadir_root
datadir_interim = interim

# locations of raw data files
# relative to $datadir_root
datadir_raw = raw

# locations of redux data files
# relative to $datadir_root
datadir_redux = redux

# locations of ADS data files
# relative to $datadir_root/$datadir_redux
datadir_ADS = ADS

# locations of orcid data files
# relative to $datadir_root/$datadir_redux
datadir_orcid = ordic

# locations of model files
# relative to $datadir_root
datadir_models = models

# locations of report files
# relative to $datadir_root
datadir_report = report


# ____________________________________________________
```

(continues on next page)

(continued from previous page)

```
[pp] # PIPELINE

# Select which steps in the data reduction pipeline must be run.

# steps 01 are mandatory

# steps 02:

# use OAC data
use_OAC_data = yes

# use IATE data
use_IATE_data = yes

# use IALP data
use_IALP_data = yes

# use GAE data
use_GAE_data = yes

# use IAFE data
use_IAFE_data = yes

# use ICATE data
use_ICATE_data = yes

# use CIC data
use_CIC_data = yes

# generate gender data
gen_gender = yes

# generate age data
gen_age = yes

# download ADS data
get_ads_data = yes

# guess orcid data
guess_orcid_data = yes

# build journals indices
build_journals_indices = yes

# generate value added publication data
build_valueadded_pub = yes


# _____________________________________________________
[run] # CONFIGURATIONS FOR EXPERIMENT AND COMPUTATIONS

# performance computing ---
```

(continues on next page)

(continued from previous page)

```
# number of jobs, to be passed to joblib.  Ignored if not run_parallel:
n_jobs = 1
# whether to run serial or parallel:
run_parallel = no


# ________________________________________________________
[out] # OUTPUT SETTINGS



# ________________________________________________________
[UX] # USER EXPERIENCE

# Show progress bars
# options: Y/N
show_progress = y

# Show messages for partial computations
# options: Y/N
verbose = y

# Return objects (N: only write to files)
# options: Y/N
interactive = n
```

The directory tree structure is defined as follows:

```
├── astrogen
│   ├── data
│   ├── dataviz
│   ├── models
│   └── sql
├── data
│   ├── external
│   │   ├── ADS
│   │   └── ORCID
│   ├── interim
│   │   └── ADS
│   ├── collect
│   └── redux
├── docs
│   └── source
│       ├── api
│       ├── img
│       └── project
├── models
├── notebooks
├── figures
└── sets
```

This structure must be used with the configuration file defaults. If a different structure is needed, the corresponding

names of the directories must be changed, of the code edited so as to ignore the parsing of the configuration file and override the default values.

Once the settings have been saved, run the pipeline:

```
cd astrogen/astrogen/data
python pipeline
```

This code generates a pickle file containing a pandas dataframe with the full dataset. An SQL data file similar to the one provided can be generated adding the following steps:

```
python clean_anonymous
python database_anonymous
```

CHAPTER

NINE

FIGURES

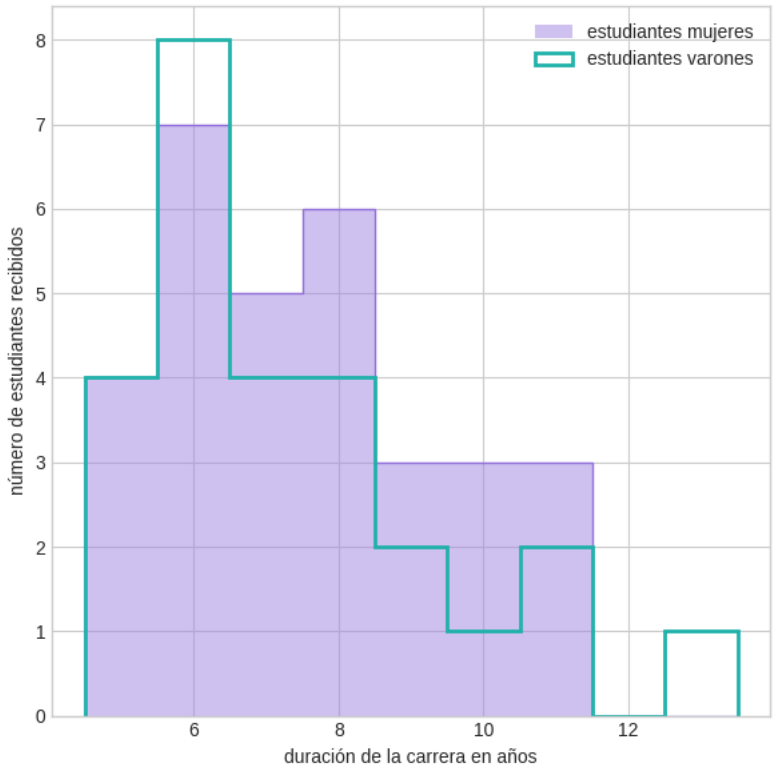
The figures to analyze the performance in the astronomy career according to genders can be easily generated with provided code.

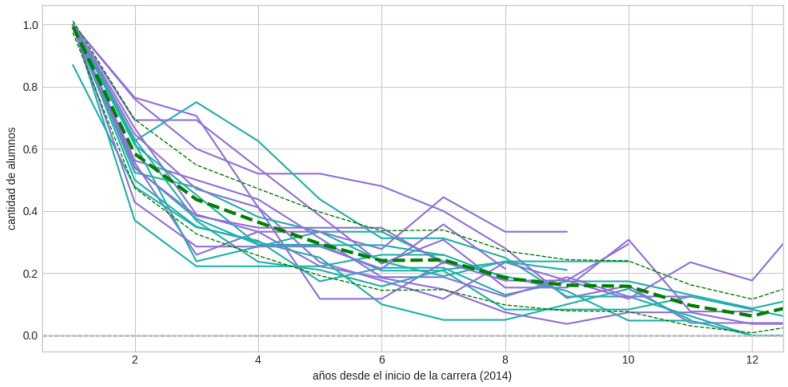
The code is located in ASTROGEN/astrogen/dataviz, and plots are saved into astrogen/figures.

Data to make these plots is in the file astrogen_DB_anonymized.db, which must be placed in astrogen/data/redux, and can be downloaded from dryad…

9.1 UNC

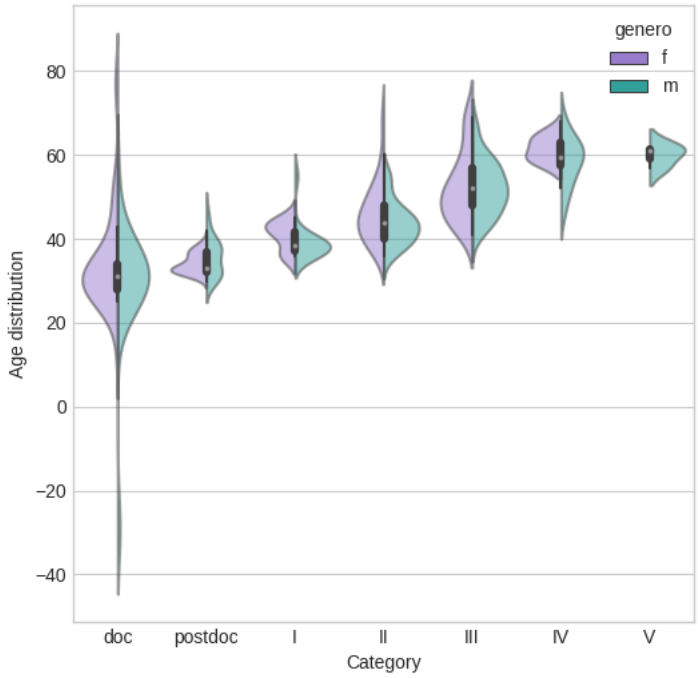
```
cd astrogen/astrogen/dataviz
python visualize_unc
```

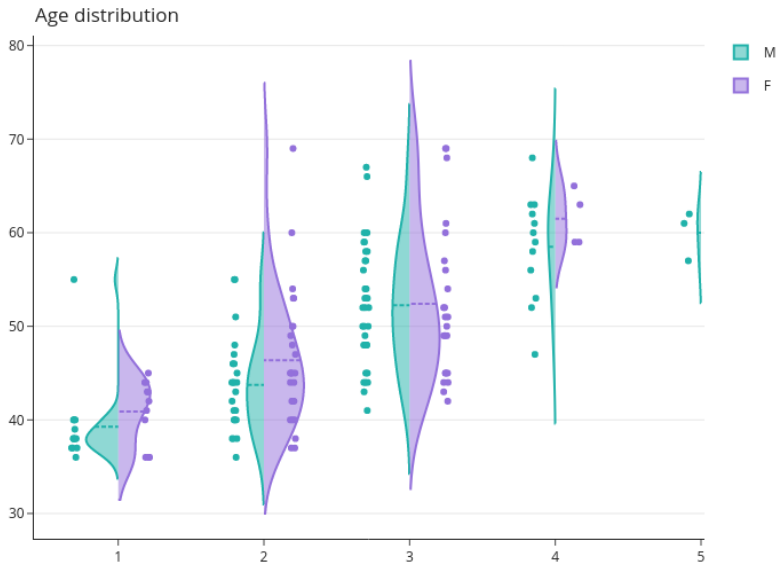




9.2 CONICET

```
cd astrogen/astrogen/dataviz
python visualize_conicet
```


genero
f
m
80
60
40
20
0
−20
−40
Age distribution
doc
postdoc
I
II
III
IV
V
Category


Age distribution
80
70
60
50
40
30
1
2
3
4
5
M
F

9.3 PUBLICATIONS

```
cd astrogen/astrogen/dataviz
python visualize_publications
```

CHAPTER

TEN

INDICES

- genindex
- modindex
- search

PYTHON MODULE INDEX

a

INDEX