Python Library for Linguistic Typology

Michael Voronov
Scientific Advisor: Boris Orekhov

Higher School of Economics

18.06.2019

Introduction

Problem:

- No Python tools for online linguistic databases queries.
- No Python tools for linguistic interactive mapping.

What exists?

- R package **lingtypology** that does both (Moroz 2017).

Why Python?

- De-facto standard language among linguists.
- A lot of scientific libraries (Pandas, SciPy etc.)
- Unicode out of the box.
- Relatively high speed.
- Versatile language.

Used Tools

- Python (Python Software Foundation 2019)
- Pandas (Augspurger et al. 2019)
- Folium (Filipe et al. 2019)
- Matplotlib (Caswell et al. 2019)
- PyGlottolog (Forkel 2019)
- SciPy (Jones, Oliphant, Peterson, et al. 2019)

Project Description

Remote Repository:

- https://github.com/OneAdder/lingtypology

Documentation:

- https://oneadder.github.io/lingtypology/

Modules:

- lingtypology.maps
- lingtypology.datasets
- lingtypology.glottolog

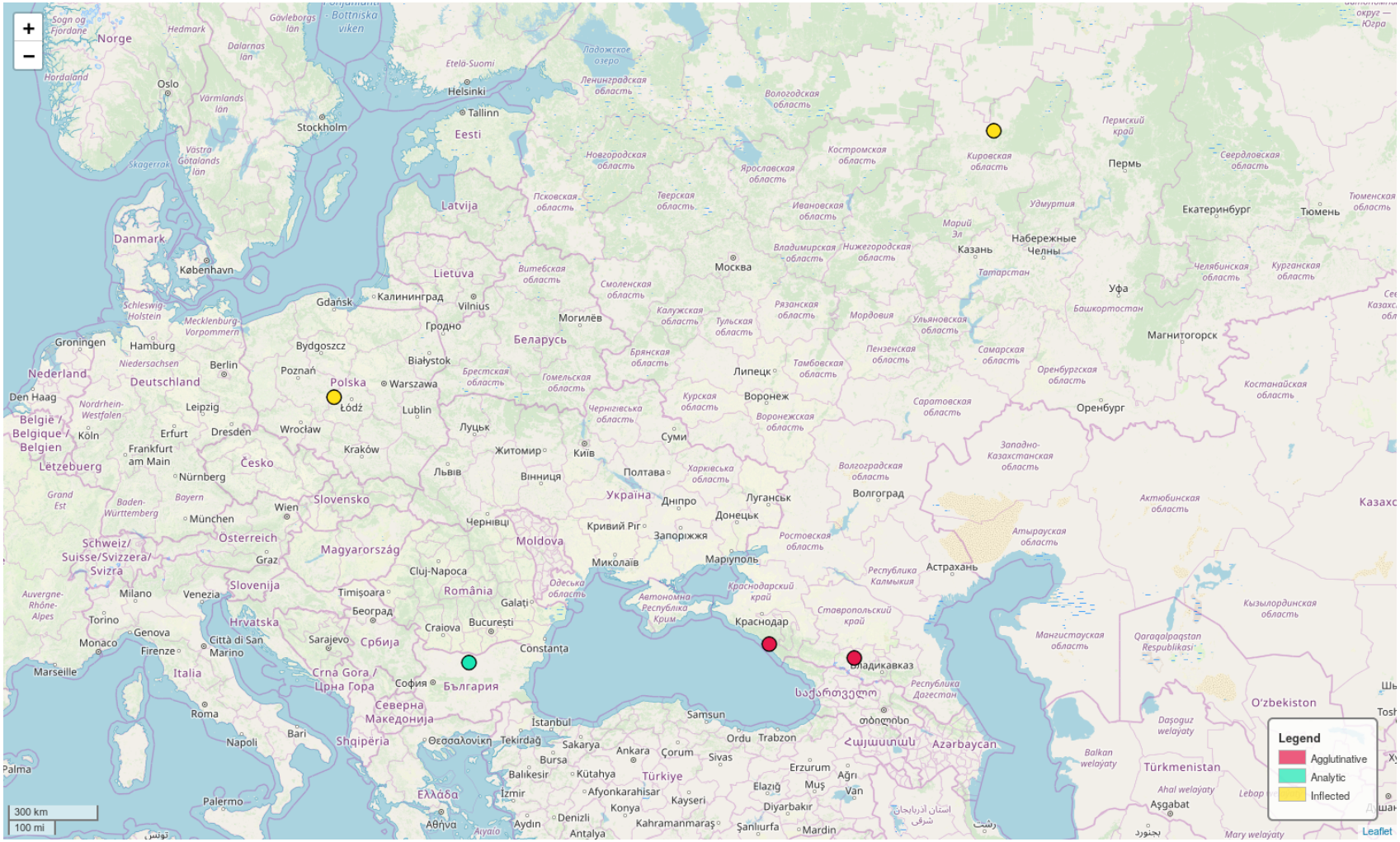
Maps

```
languages =('Romanian', 'Afrikaans', 'Tlingit', 'Japanese')
m =lingtypology.LingMap(languages)
m.create_map()
```

Maps

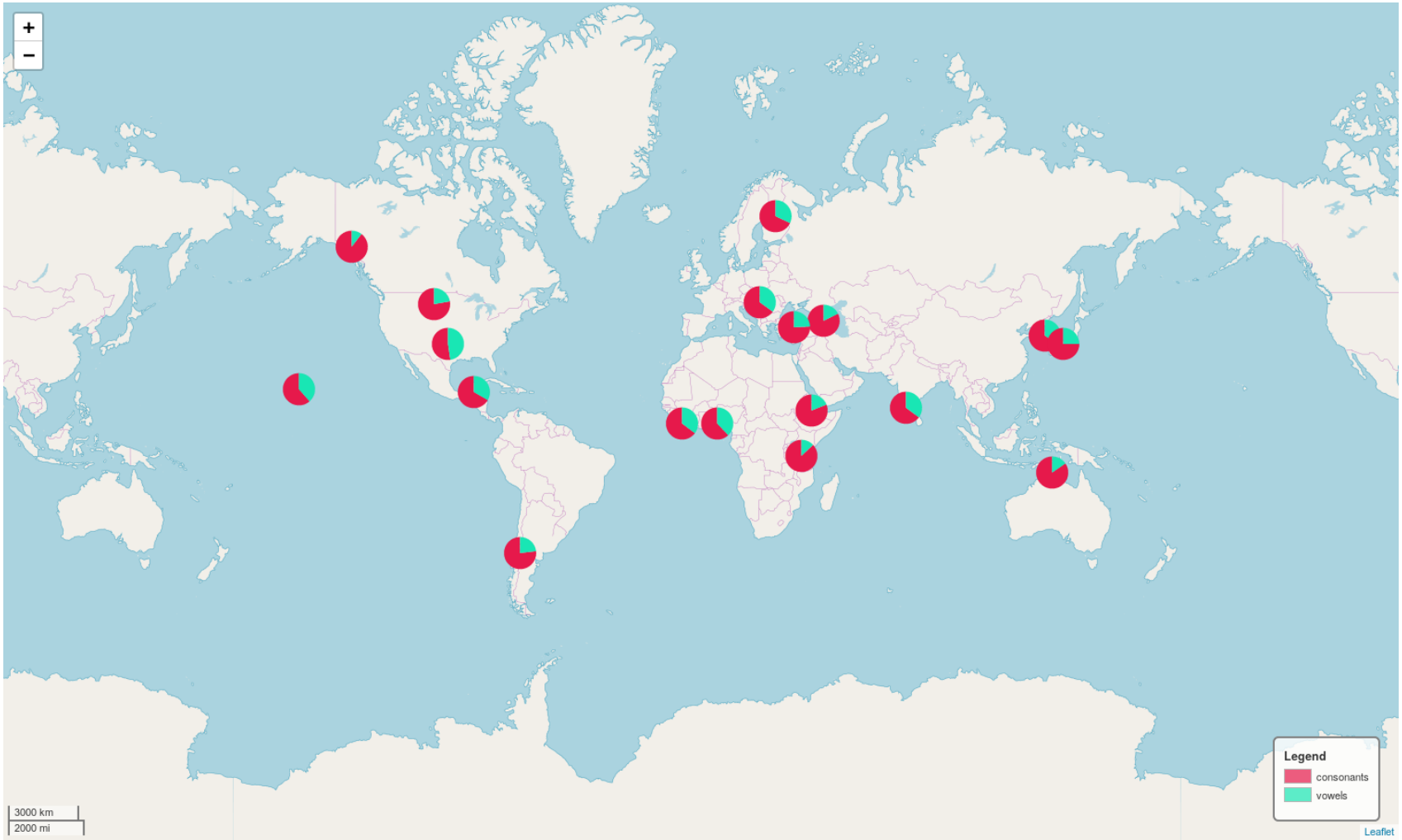
```
languages =[
      ”Adyghe”, ”Kabardian”, ”Polish”,
      ”Russian”, ”Bulgarian”
]
features =[
      ”Agglutinative”, ”Agglutinative”, ”Inflected”,
      ”Inflected”, ”Analytic”
]
m =lingtypology.LingMap(languages)
m.add_features(features)
m.create_map()
```

Maps


Legend
Agglutinative
Analytic
Inflected
300 km
100 mi
Leaflet

Maps

```
m =lingtypology.LingMap(data.language)
m.add_minicharts(data.consonants, data.vowels)
m.create_map()
```



Databases

- WALS: The World Atlas of Language Structures (Dryer and Haspelmath 2013).
- Autotyp: an international network of typological linguistic databases (Bickel et al. 2017).
- AfBo: A world-wide survey of affix borrowing (Seifart 2013).
- SAILS: The South American Indigenous Language Structures (Muysken et al. 2016).
- PHOIBLE: ... is a repository of cross-linguistic phonological inventory data (Moran and McCloy 2019).

WALS

```
w =lingtypology.datasets.Wals('1a')
w.get_df().head(10)
```

	wals_code	language	genus	family	coordinates	_1A_area	_1A	_1A_num	_1A_desc
0	kiw	Kiwai (Southern)	Kiwaian	Kiwaian	(-8.0, 143.5)	Phonology	1. Small	1	Small
1	xoo	!Xóõ	Tu	Tu	(-24.0, 21.5)	Phonology	5. Large	5	Large
2	ani	//Ani	Khoe-Kwadi	Khoe-Kwadi	(-18.9166666667, 21.9166666667)	Phonology	5. Large	5	Large
3	abi	Abipón	South Guaicuruan	Guaicuruan	(-29.0, -61.0)	Phonology	2. Moderately small	2	Moderately small
4	abk	Abkhaz	Northwest Caucasian	Northwest Caucasian	(43.0833333333, 41.0)	Phonology	5. Large	5	Large
5	acm	Achumawi	Palaihnihan	Hokan	(41.5, -121.0)	Phonology	2. Moderately small	2	Moderately small
6	ach	Aché	Tupi-Guaraní	Tupian	(-25.25, -55.1666666667)	Phonology	1. Small	1	Small
7	aco	Acoma	Keresan	Keresan	(34.9166666667, -107.583333333)	Phonology	5. Large	5	Large
8	adz	Adzera	Oceanic	Austronesian	(-6.25, 146.25)	Phonology	2. Moderately small	2	Moderately small
9	agh	Aghem	Bantoid	Niger-Congo	(6.666666666669999, 10.0)	Phonology	3. Average	3	Average

WALS

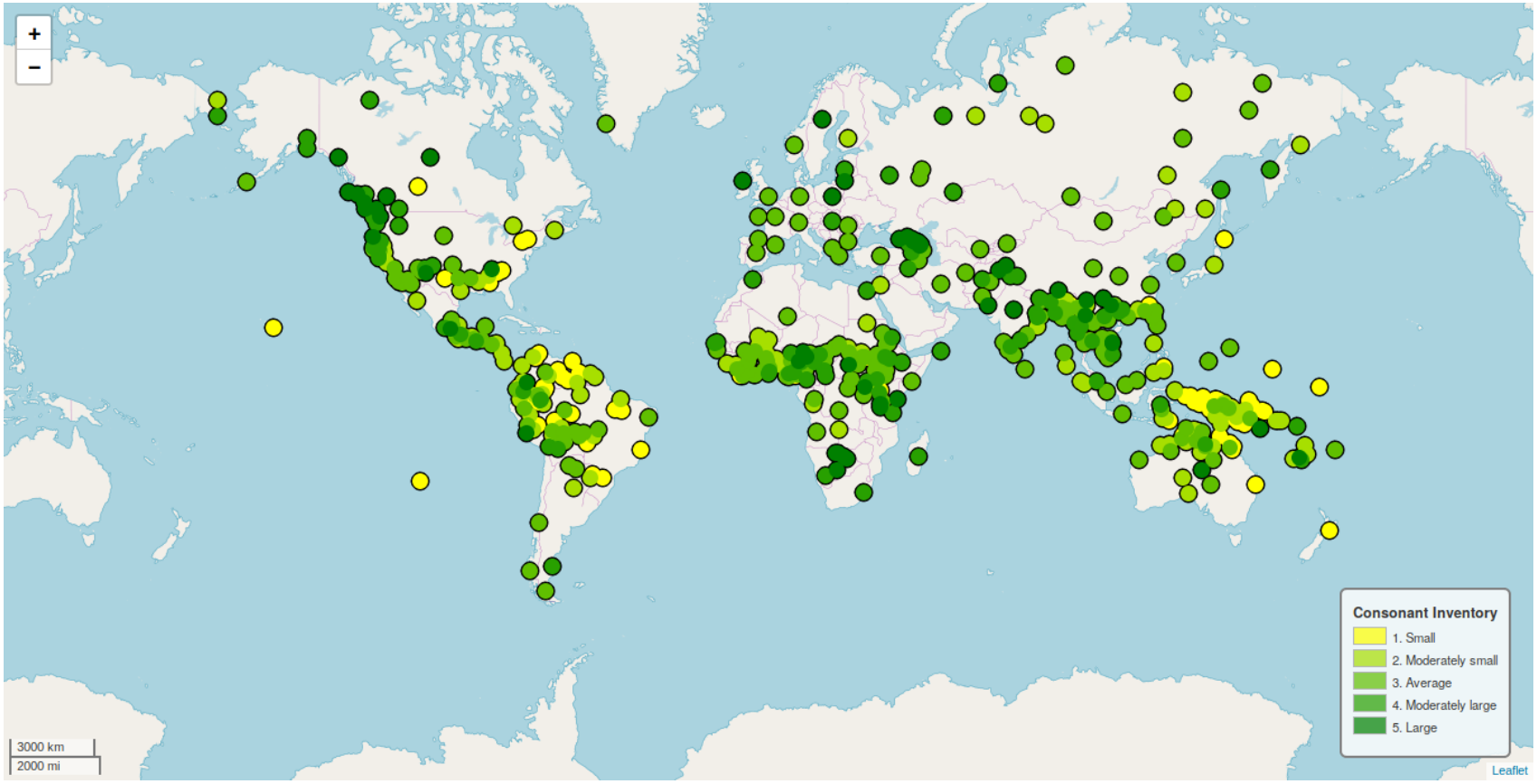
```
w =lingtypology.datasets.Wals('1a', '2a')
w.get_df().head()
```

	language	...	_1A	...	_2A	...
0	Kiwai (Southern)	...	1. Small	...	2. Average (5-6)	...
1	!Xóõ	...	5. Large	...	2. Average (5-6)	...
2	//Ani	...	5. Large	...	2. Average (5-6)	...
3	Abipón	...	2. Moderately small	...	2. Average (5-6)	...
4	Abkhaz	...	5. Large	...	1. Small (2-4)	...

Examples: WALS Features

```
wals_page =lingtypology.datasets.Wals('1a').get_df()
m =lingtypology.LingMap(wals_page.language)
m.add_custom_coordinates(wals_page.coordinates)
m.add_features(
    wals_page._1A,
    colors=lingtypology.gradient(5, 'yellow', 'green')
)
m.legend_title ='Consonant Inventory'
m.create_map()
```

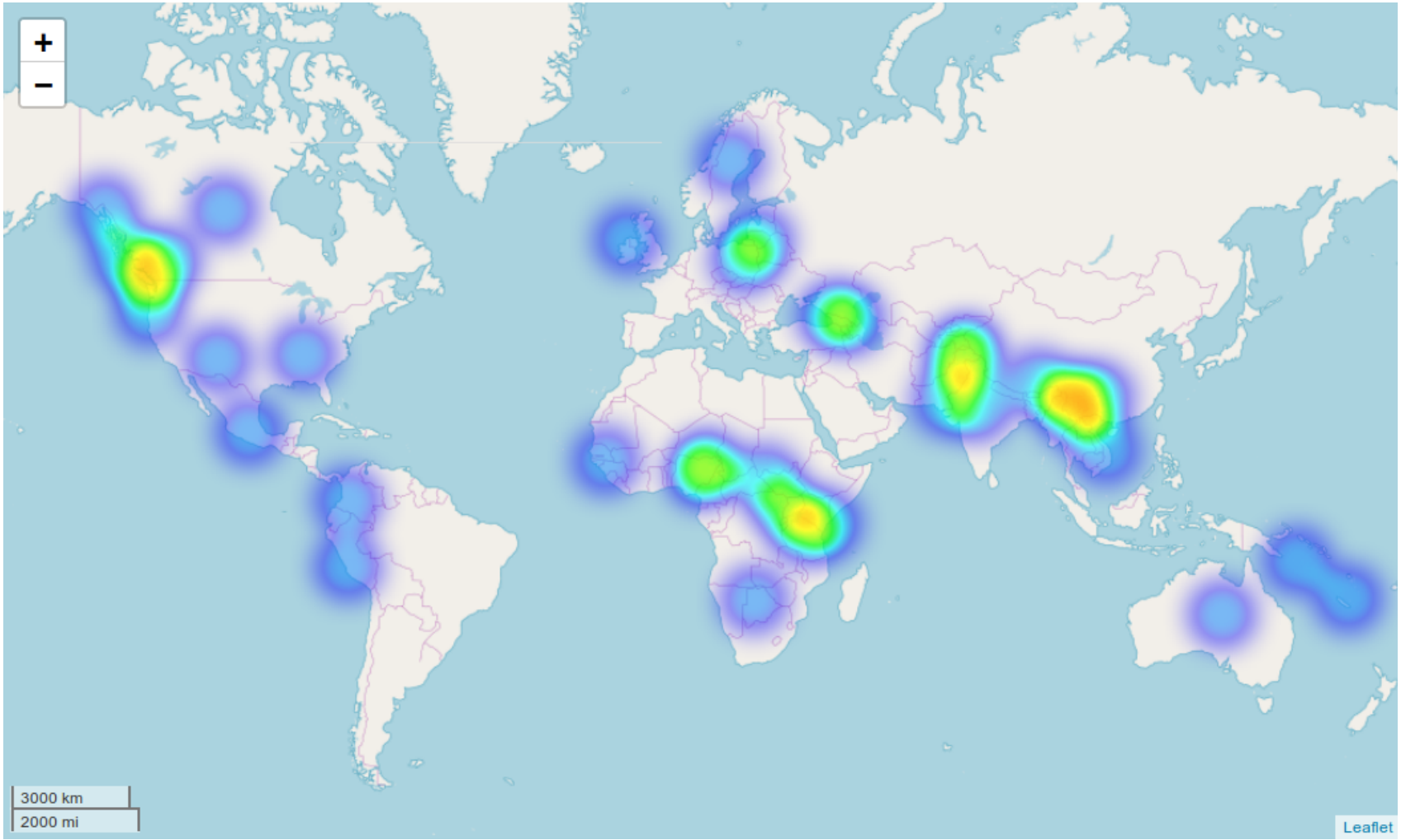
Examples: WALS Features



Examples: WALS Heatmap

```
wals =lingtypology.datasets.Wals('1A')
data =wals.get_df()
m =lingtypology.LingMap()
m. add_heatmap(data[data._1A_desc =='Large'].coordinates)
m.create_map()
```

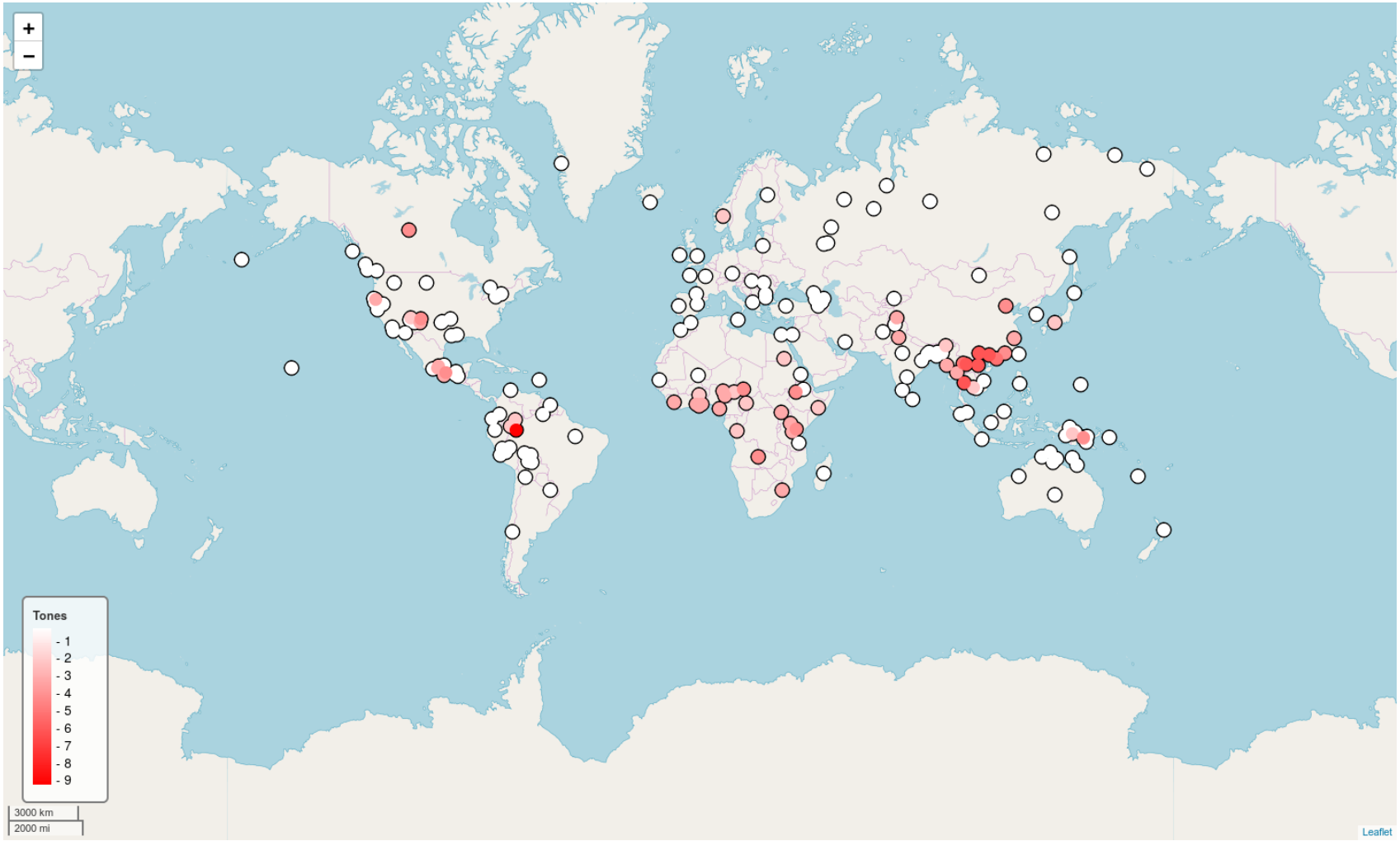
Examples: WALS Heatmap



Examples: PHOIBLE Tones

```
p =lingtypology.datasets.Phoible(subset='SPA')
df =p.get_df(strip_na=['tones'])
m =lingtypology.LingMap(df.language)
m.add_features(df.tones, numeric=True)
m.colormap_colors =('white', 'red')
m.legend_title ='Tones'
m.legend_position ='bottomleft'
m.create_map()
```

Examples: PHOIBLE Tones



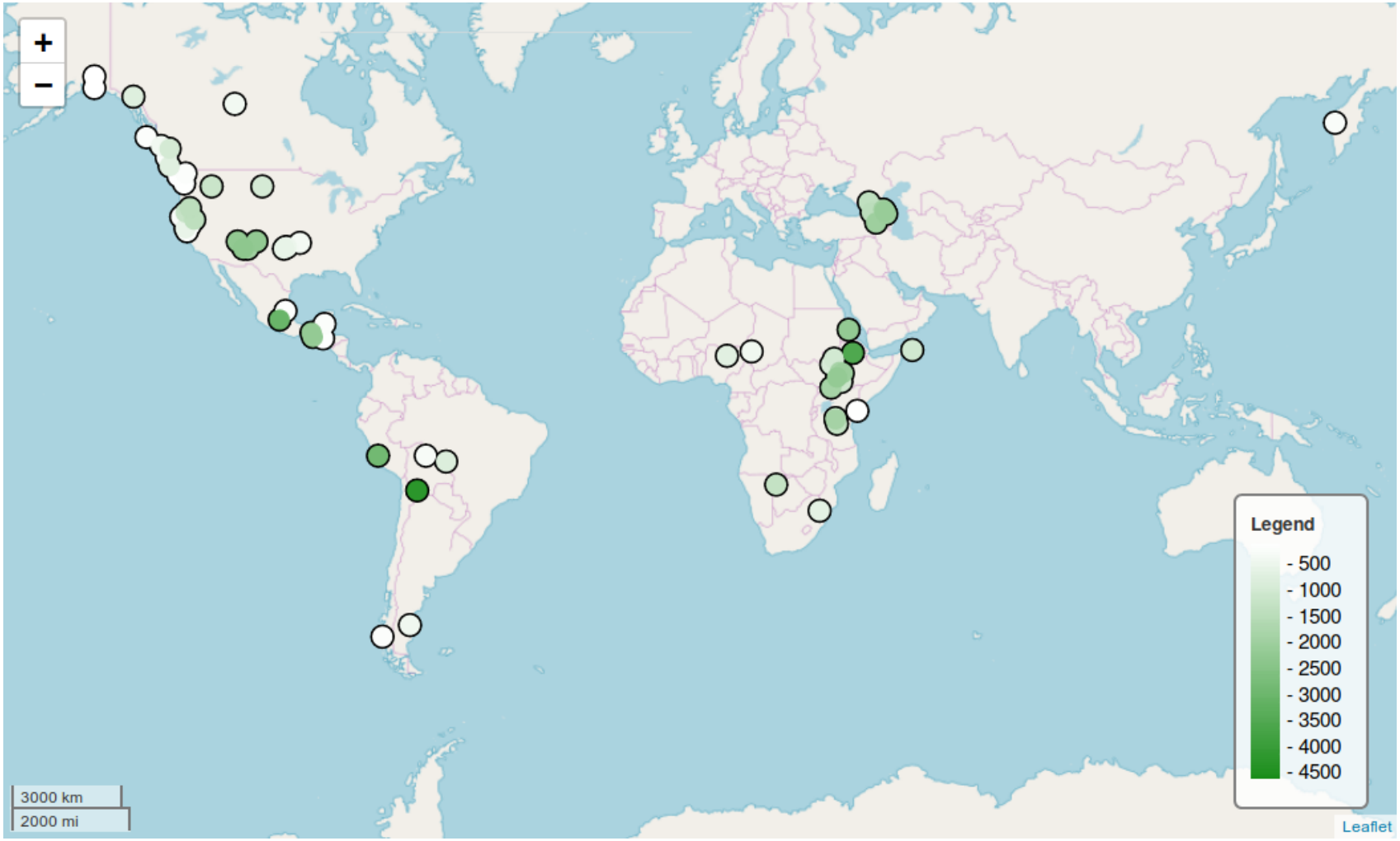
Verification of Statistical Studies in the Context of Reproducibility

- Article that demonstrates relationship between presence of ejectives and high elevation based on WALS data (Everett 2013).
- Reproduce on PHOIBLE data.

Verification of Statistical Studies in the Context of Reproducibility

```
upsid =lingtypology.datasets.Phoible(
    subset='UPSID',
    aggregated=False
).get_df()
amount_of_ejectives =upsid[
    upsid.raisedLarynxEjective =='+'
].groupby('Glottocode').size()
languages =[
    lingtypology.glottolog.get_by_glot_id(glot_id) \
    for glot_id in amount_of_ejectives.index
]
upsid_ejectives =pandas.DataFrame({
    'language': languages,
    'ejectives': amount_of_ejectives,
    'elevation': lingtypology.get_elevations(languages),
})
m =lingtypology.LingMap(upsid_ejectives.language)
m.add_features(upsid_ejectives.elevation, numeric=True)
m.create_map()
```

Verification of Statistical Studies in the Context of Reproducibility



Verification of Statistical Studies in the Context of Reproducibility

PHOIBLE datasets:

- SAPHON: South American Phonological Inventory Database (Lev, Stark, and Chang 2012).
- AA: Alphabets of Africa (Chanard 2006).
- GM: ‘Christopher Green and Steven Moran extracted phonological inventories from secondary sources including grammars and phonological descriptions with the goal of attaining pan-Africa coverage‘ (Moran, McCloy, and Wright 2014).

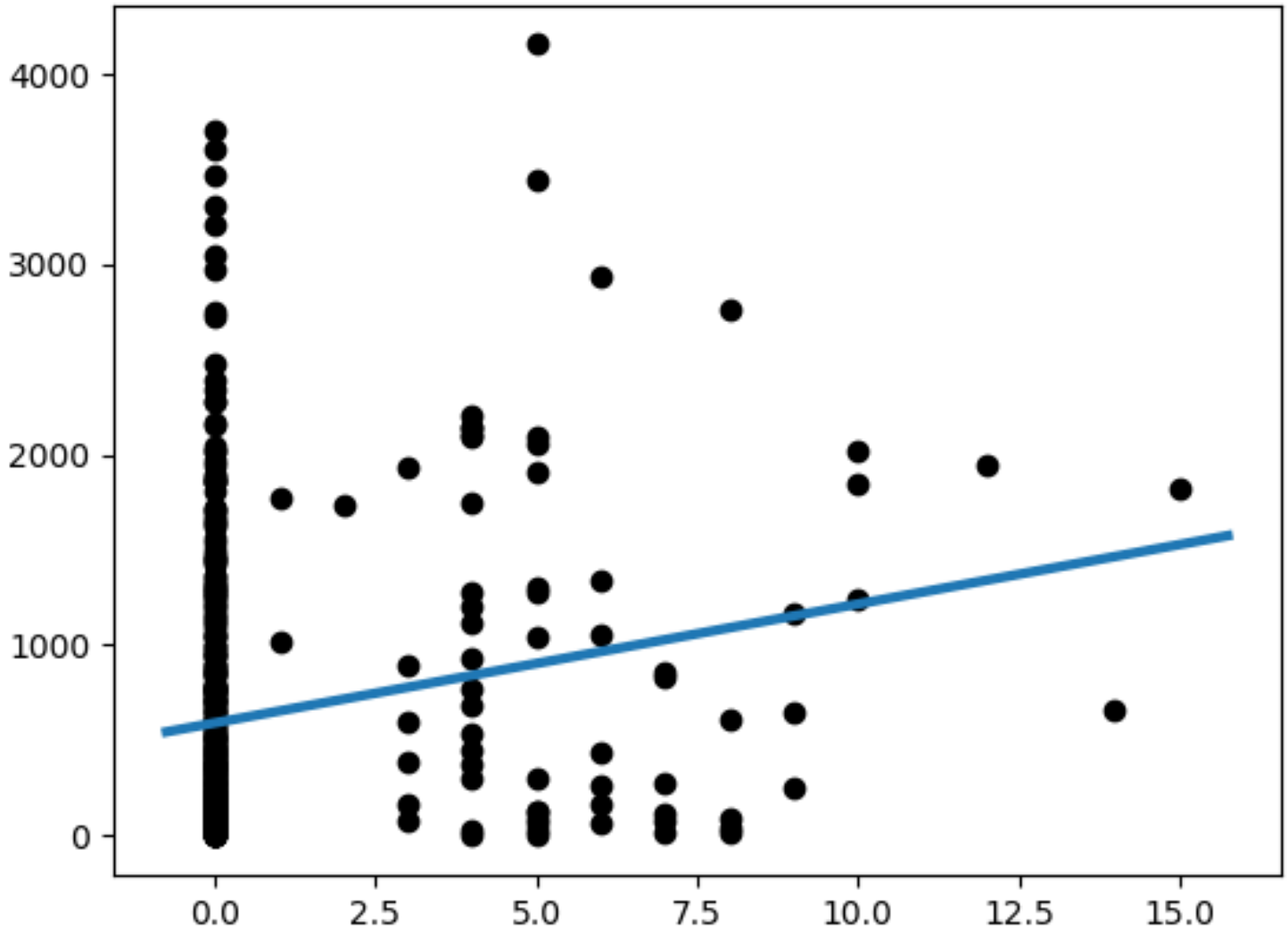
Verification of Statistical Studies in the Context of Reproducibility

- PH: ‘Christopher Green and Steven Moran extracted phonological inventories from secondary sources including grammars and phonological descriptions with the goal of attaining pan-Africa coverage‘ (Moran, McCloy, and Wright 2014).
- RA: Common Linguistic Features in Indian Languages: Phoentics (Ramaswami 1999).
- SPA: Stanford Phonology Archive (Crothers et al. 1979).
- UPSID: UCLA Phonological Segment Inventory Database (Maddieson and Precoda 1990).

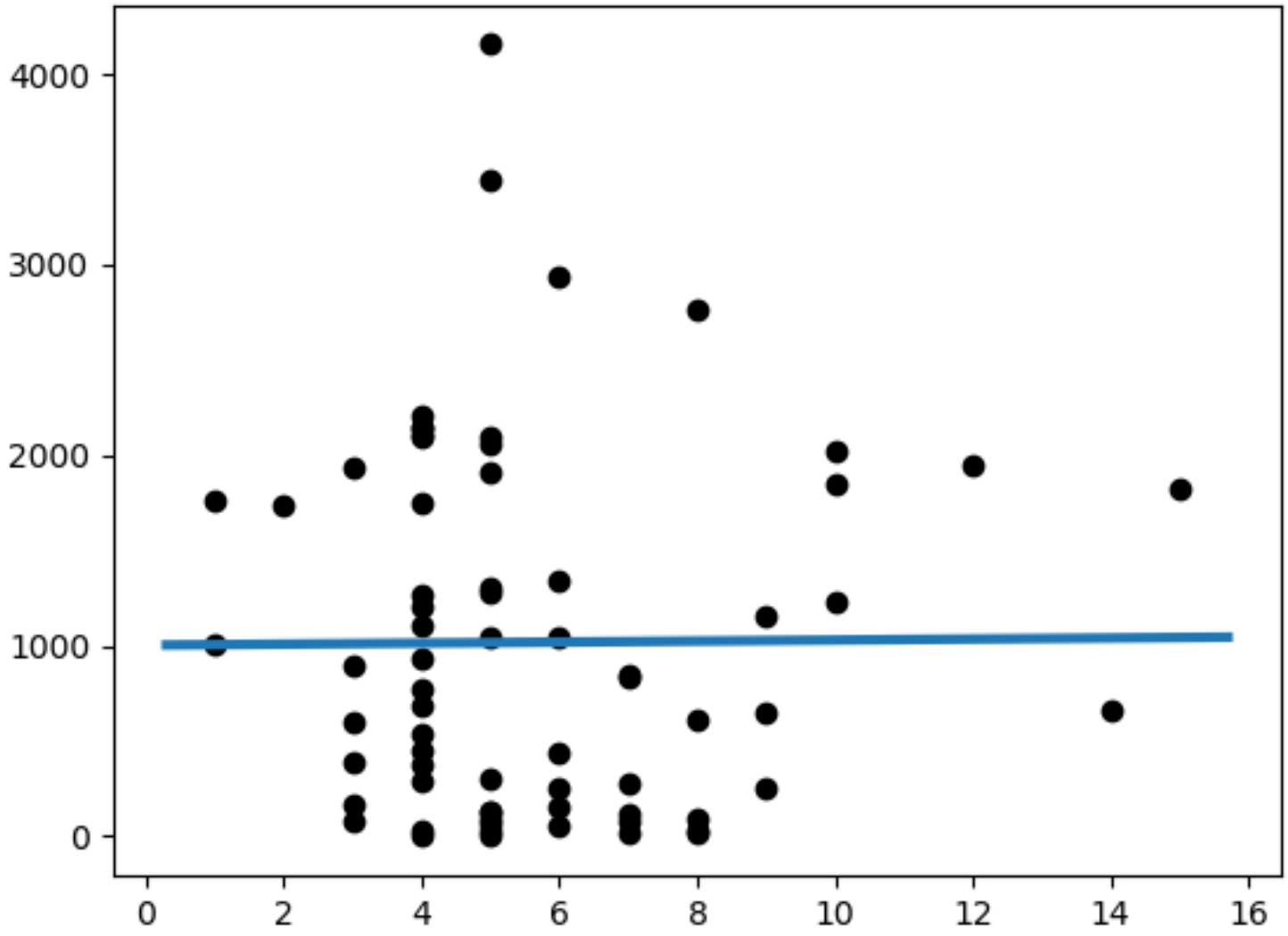
Verification of Statistical Studies in the Context of Reproducibility

	Dataset	Regression (with ejectives only)	Regression (all languages)	Chi2 Test
0	UPSID	0.95055	0.00004	0.00003
1	SPA	0.47553	0.00001	0.00018
2	PH	0.73152	0.39245	0.16019
3	GM	0.03858	0.00000	0.00000
4	SAPHON	0.018874	0.00000	0.00038

Verification of Statistical Studies in the Context of Reproducibilitv



Verification of Statistical Studies in the Context of Reproducibilitv



Verification of Statistical Studies in the Context of Reproducibility

Results:

- True: share of languages with ejectives is higher if the elevation is more than 1500m (verified on PHOIBLE data).
- Not true: the higher the language, the more ejectives there are.

PHOIBLE and Elevation

	Dataset	short	long	delayedRelease	...
0	UPSID	0.7304	0.6205	0.6106	...
1	SPA	0.4974	0.8311	0.4335	...
2	GM	0.6587	0.0070	0.8435	...
3	RA	0.0826	0.1125	nan	...
5	AA	NaN	0.7559	nan	...
6	PH	NaN	0.2549	0.9051	...
7	SAPHON	NaN	0.0287	0.4856	...
4	Median	0.578074	0.254949	0.610642	...

Full Table:
https://github.com/OneAdder/lingtypology_research/blob/master/PHOIBLE:%20Quantitative%20Research.ipynb

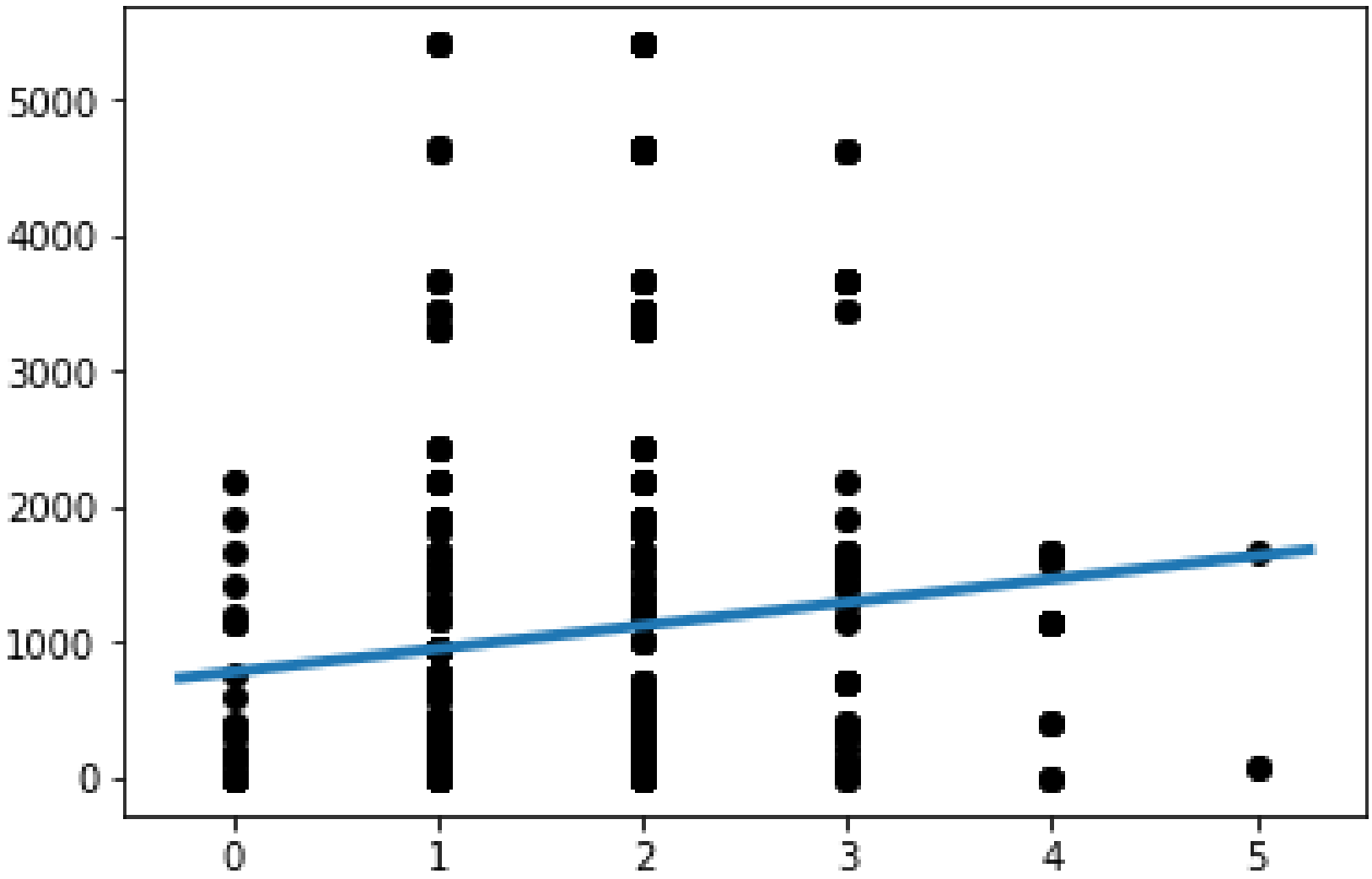
Autotyp and Elevation

- ‘Exponence: number of categories that are expressed in the same marker‘.
- ‘Rough approximation of the size of the possessum category in terms of the number of semantic classes covered‘.
- ‘Number of separately marked inflectional categories (including agreement) in position ”post” of the verb‘.
- ‘Number of morpheme types included in a phonologically or grammatically coherent suffix domain‘.

Autotyp and Elevation

Feature	Subfeature	P-value
Grammatical_markers	Exponence.n	0.00000000
NP_structure	NPHeadSemClassSize.n	0.01766784
VInfl_counts_per_position	VInflCatAndAgrPost.n	0.02895302
Word_domains	MphmTypesInCohSuffixDomain.n	0.00196901

Autotyp and Elevation



WALS: Implicative Universaliae

feature	_10A_desc	_25B_desc	_39B_desc	_47A_desc	...
_10A_desc	1.00000	0.99444	nan	0.63296	...
_25B_desc	0.90442	1.00000	nan	0.96609	...
_39B_desc	1.00000	nan	1.00000	0.66501	...
_47A_desc	0.82120	0.84267	0.66501	1.00000	...
...	...	...	...	...	...

Full table:

https://github.com/OneAdder/lingtypology_research/blob/master/WALS:%20Quantitative%20Research.ipynb

Conclusion

- LingTypology: a Python tool for linguistic typology
  - Repository: https://github.com/OneAdder/lingtypology
  - Documentation: https://oneadder.github.io/lingtypology/
  - PyPI: https://pypi.org/project/lingtypology/
- Demonstrative Studies
  - Simplicity
  - Reproducibility
  - Visualization

References I


Augspurger, Tom et al. (2019). *pandas: powerful Python data analysis toolkit*. URL: http://pandas.pydata.org/pandas-docs/stable.


Bickel, Balthasar et al. (2017). *The AUTOTYP typological databases. Version 0.1.0*. URL: https://github.com/autotyp/autotyp-data/tree/0.1.0.


Caswell, Thomas A et al. (May 2019). *matplotlib/matplotlib v3.1.0*. DOI: 10.5281/zenodo.2893252. URL: https://doi.org/10.5281/zenodo.2893252.


Chanard, C. (2006). *Systèmes Alphabétiques Des Langues Africaines*. URL: http://sumale.vjf.cnrs.fr/phono/.


Crothers, John H. et al. (1979). "Handbook of Phonological Data From a Sample of the World's Languages: A Report of the Stanford Phonology Archive". In:


Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. URL: https://wals.info/.


Everett, Caleb (2013). "Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives". In: DOI: 10.1371/journal.pone.0065275. URL: https://doi.org/10.1371/journal.pone.0065275.


Filipe et al. (May 2019). *python-visualization/folium: v0.9.1*. DOI: 10.5281/zenodo.3229045. URL: https://doi.org/10.5281/zenodo.3229045.


Forkel, Robert (Apr. 2019). *clld/pyglottolog: Glottolog API*. DOI: 10.5281/zenodo.2620250. URL: https://doi.org/10.5281/zenodo.2620250.


Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2019). *SciPy: Open source scientific tools for Python*. URL: http://www.scipy.org/.


Lev, Michael, Tammy Stark, and Will Chang (2012). *South American Phonological Inventory Database*. URL: http://linguistics.berkeley.edu/%20saphon/en/.

References II


Maddieson, Ian and Kristin Precoda (1990). "Updating UPSID". In: *UCLA Working Papers in Phonetics*. Vol. 74. Department of Linguistics, UCLA, pp. 104–111.


Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. URL: https://phoible.org/.


Moran, Steven, Daniel McCloy, and Richard Wright, eds. (2014). *PHOIBLE Online*. URL: http://phoible.org/.


Moroz, George (2017). *lingtypology: easy mapping for Linguistic Typology*. DOI: 10.5281/zenodo.1289471. URL: https://CRAN.R-project.org/package=lingtypology.


Muysken, Pieter et al. (2016). *South American Indigenous Language Structures (SAILS) Online*. URL: http://sails.clld.org.


Python Software Foundation (2019). *The Python Language Reference*. URL: https://docs.python.org/3.7/reference/.


Ramaswami, N. (1999). *Common Linguistic Features in Indian Languages: Phonetics*. Central Institute of Indian Languages.


Seifart, Frank, ed. (2013). *AfBo: A world-wide survey of affix borrowing*. URL: https://afbo.info/.