Geographic origin detection from commit data in open source projects

Ali Mustapha, **Davide Rossi**, Stefano Zacchiroli

OpenSym & OSS 2023 Gather Together

Problem statement

- Characterize the geographic diversity of public code contributors.
- *“Where do people committing to open source projects come from?”*

Problem statement

- Characterize the geographic diversity of public code contributors.
- *“Where do people committing to open source projects come from?”*
- Will you find a response here? No, here we try to address a specific sub-problem. Moreover this is a WIP.

Previous work

- Gonzalez-Barahona, J.M., Robles, G., Andradas-Izquierdo, R. and Ghosh, R.A., 2008. Geographic origin of libre software developers. Information Economics and Policy, 20(4), pp.356-363.
- Rossi, D. and Zacchiroli, S., 2022. Geographic diversity in public code contributions: an exploratory large-scale study over 50 years. In Proceedings of the 19th International Conference on Mining Software Repositories, pp. 80-85.

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

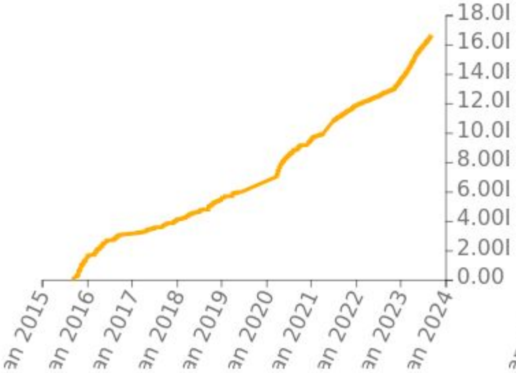
Bitbucket 2,396,928 origins	56,983 origins	git 22,544 origins
22,775 origins	debian 133,968 origins	37,695 origins
GitHub 195,350,961 origins	gitiles 9,790 origins	GitLab 4,110,503 origins
git 1,222 origins	Gogs 122 origins	GO 491,291 origins
GNU 354 origins	heptapod 1,159 origins	launchpad 488,898 origins
Maven 273,862 origins	npm 3,384,650 origins	4,839 origins
Packagist The PHP Package Repository 189,644 origins	fedora PAGURE 67,585 origins	Phabricator 212 origins
pub.dev 44,800 origins	python Package Index 476,676 origins	SOURCEFORGE 380,795 origins

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Source files

16,588,464,304

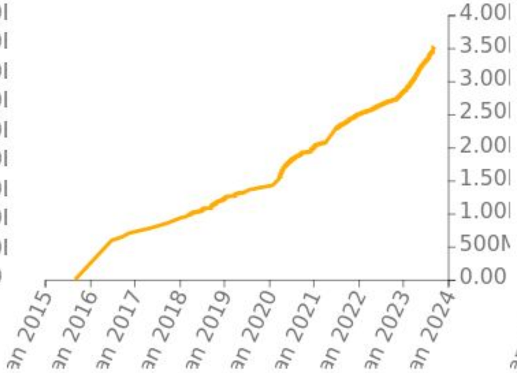


Directories

13,379,865,911

Commits

3,514,208,501

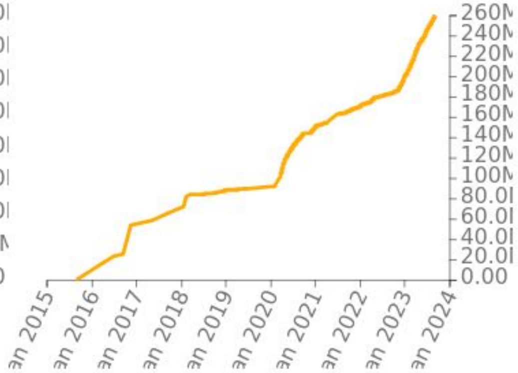


Authors

63,262,027

Projects

258,904,074



Releases

72,574,928

Dataset

- Revisions (commits) and their authors
- From 1-1-1970 to 30-5-2023
- 240 (est.) million projects
- 3 billion commits
- 43 million authors
- Sourced from different VC systems, only common metadata are usable

Features

- Commit data
  - Name
  - Email
  - Commit date, with time zone offset
  - Commit message

Features

- Commit data
    - Name (self declared. Yes we do have ‘Jabba the Hutt’)
    - Email
    - Commit date, with time zone offset
    - Commit message

Features

- Commit data
    - Name (self declared. Yes we do have ‘Jabba the Hutt’. 4 of them)
    - Email
    - Commit date, with time zone offset
    - Commit message

Features

- Commit data
  - Name (self declared. Yes we do have ‘Jabba the Hutt’. 4 of them)
  - Email (unverified; uneven adoption of national ccTLD)
  - Commit date, with time zone offset
  - Commit message

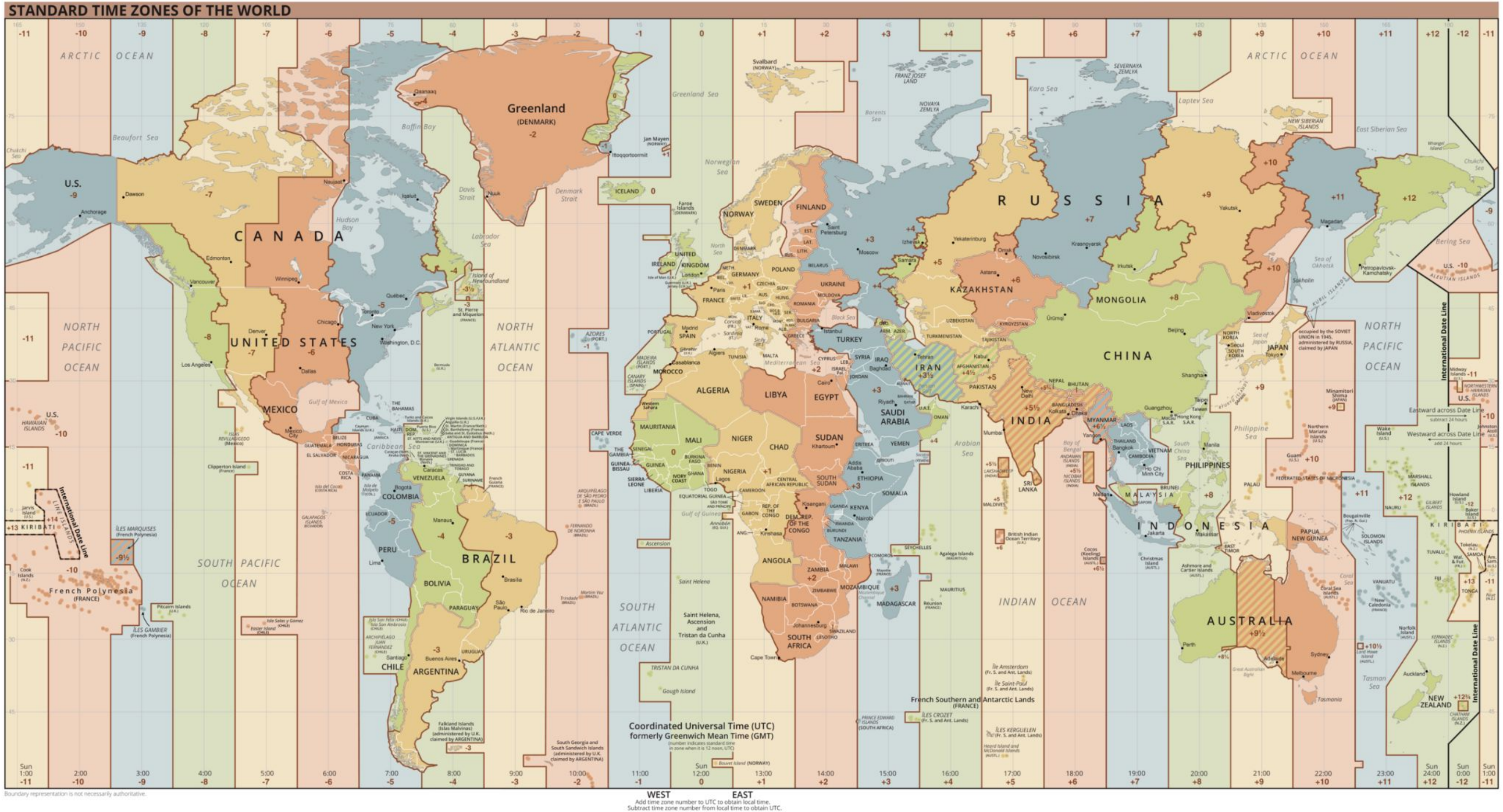
Features

- Commit data
  - Name (self declared. Yes we do have ‘Jabba the Hutt’. 4 of them)
  - Email (unverified; uneven adoption of national ccTLD)
  - Commit date, with time zone offset
  - Commit message (most usually in English)

Possible approaches

- Use offsets
- Use names
  - Dictionary-based approach
  - ML (NN) approach
- Use names and offsets

The time zones



The UN geoscheme

The zones

Russia
Northern America
Europe
Western
and Central Asia
China
Eastern Asia
Southern
Asia
Africa
South-eastern Asia
Latin America and
the Caribbean
Australia, New Zealand and
Pacific

Possible approaches

- Use offsets
- **Use names**
  - Dictionary-based approach
  - ML (NN) approach
- Use names and offsets

A few words on names and nationalities

While the idea of using persons’ names to detect the part of the world they come from seems reasonable, we should bear in mind a name is a cultural product and a nation is a political fabrication. Nations do not necessarily align with cultures, which implies that using names to detect nationalities poses various subtitle challenges.

Possible approaches

- Use offsets
- **Use names**
  - Dictionary-based approach - we need a **dictionary**
  - ML (NN) approach - we need a **training set**
- Use names and offsets

Possible approaches

- Use offsets
- **Use names**
    - Dictionary-based approach - we need a **dictionary**
    - ML (NN) approach - we need a **training set**
- Use names and offsets
- In all cases we also need a **labeled dataset** for validation

Dictionary

Ideal data: most used forenames/surnames and their frequency for each existing country.

Is this publicly available? No.

Can this be extracted from public sources? Brief answer: no.

Training set

- Is there a **large enough** publicly available labeled dataset? No.
- Our solution: synthetic names.
  - Use the name dictionary to generate full names randomly mixing forenames and surnames.
  - If datetime with offset is needed, randomly generate that too.

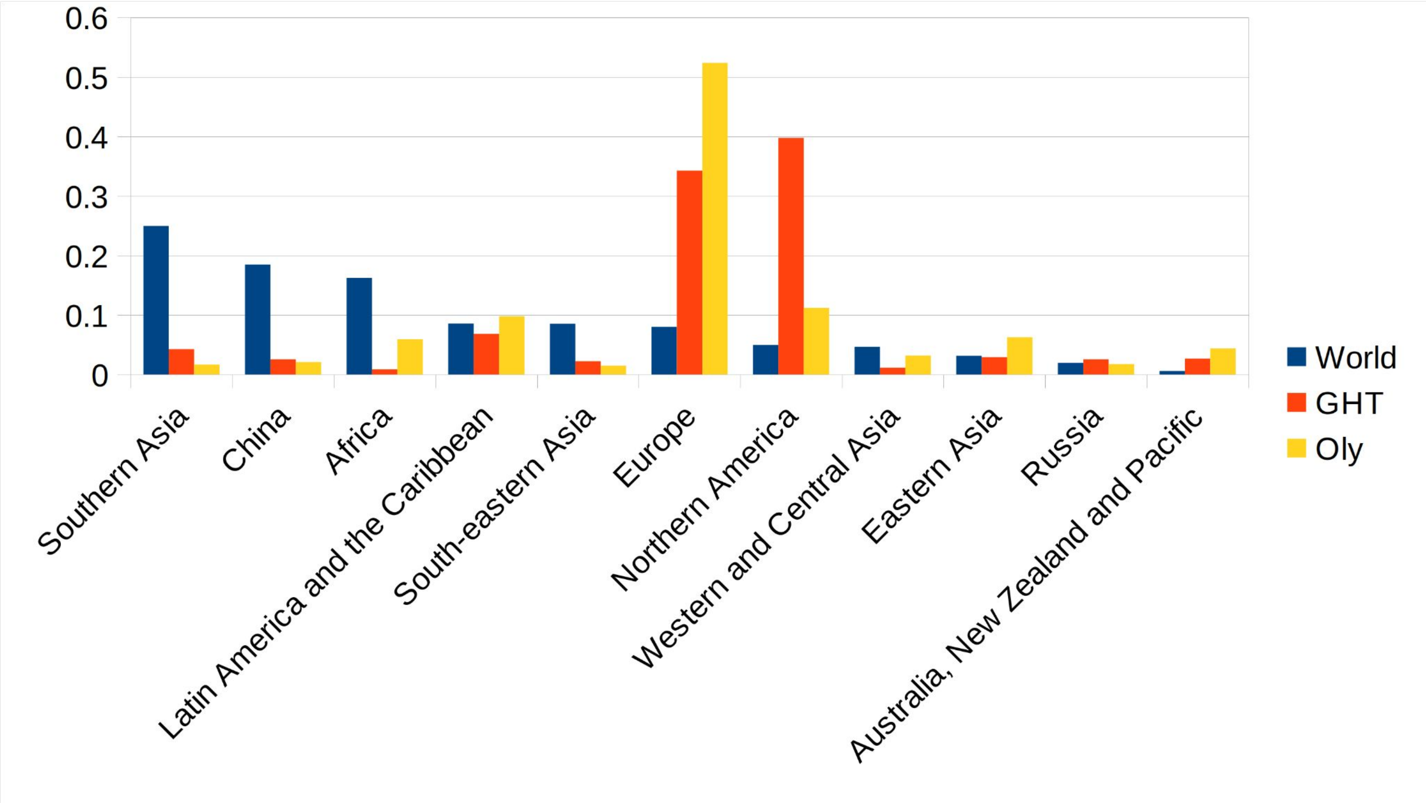
The NN

- Simple CNN with bi-gram grouping and an embedding layer.
- Trained with a synthetic dataset using a quantile-based supersampling approach: data stratification is performed on the basis of names’ frequencies and of places’ population.
- One network using the TZ offset as an input feature.
- Multiple networks, one for each TZ offset.

Validation dataset

- “Generic”: data from the Olympic games.
- “Domain specific”: data extracted from GHT. GHT provides details about Github hosted projects including commits data linked to user profiles. User profiles include a *location* field.

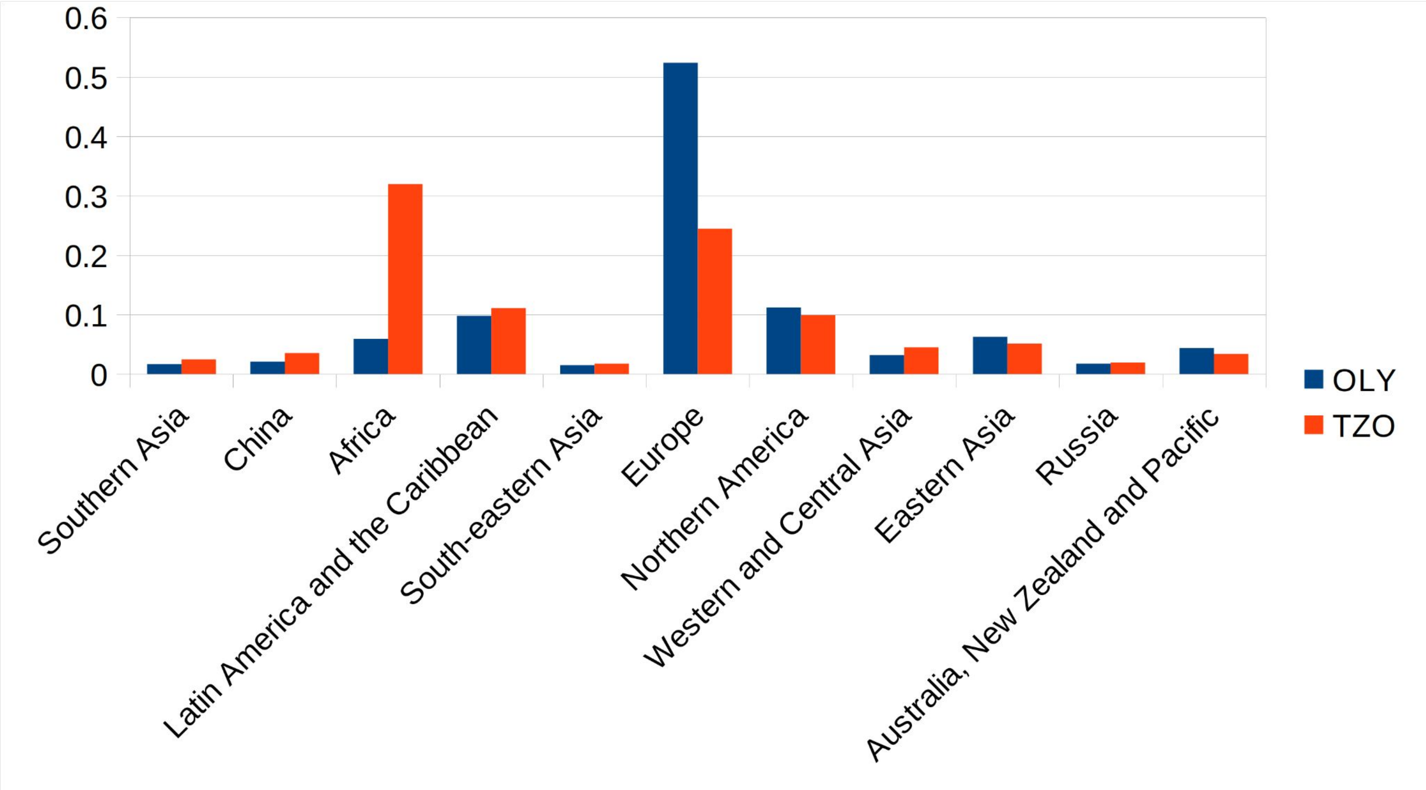
Comparing the datasets



Olympics: TZO offset only

	precision	recall	f1-score	support
Europe	0.928	0.433	0.591	66408
Northern America	0.695	0.615	0.653	14194
Latin America and the Caribbean	0.612	0.693	0.650	12410
Eastern Asia	0.962	0.789	0.867	7931
Africa	0.115	0.622	0.194	7485
Australia, New Zealand and Pacific	0.984	0.763	0.859	5529
Western and Central Asia	0.215	0.304	0.252	4013
China	0.509	0.858	0.639	2653
Russia	0.197	0.216	0.206	2203
Southern Asia	0.646	0.958	0.772	2090
South-eastern Asia	0.498	0.584	0.537	1887
accuracy	0.539	0.539	0.539	0
macro avg	0.578	0.621	0.565	126803
weighted avg	0.772	0.539	0.595	126803

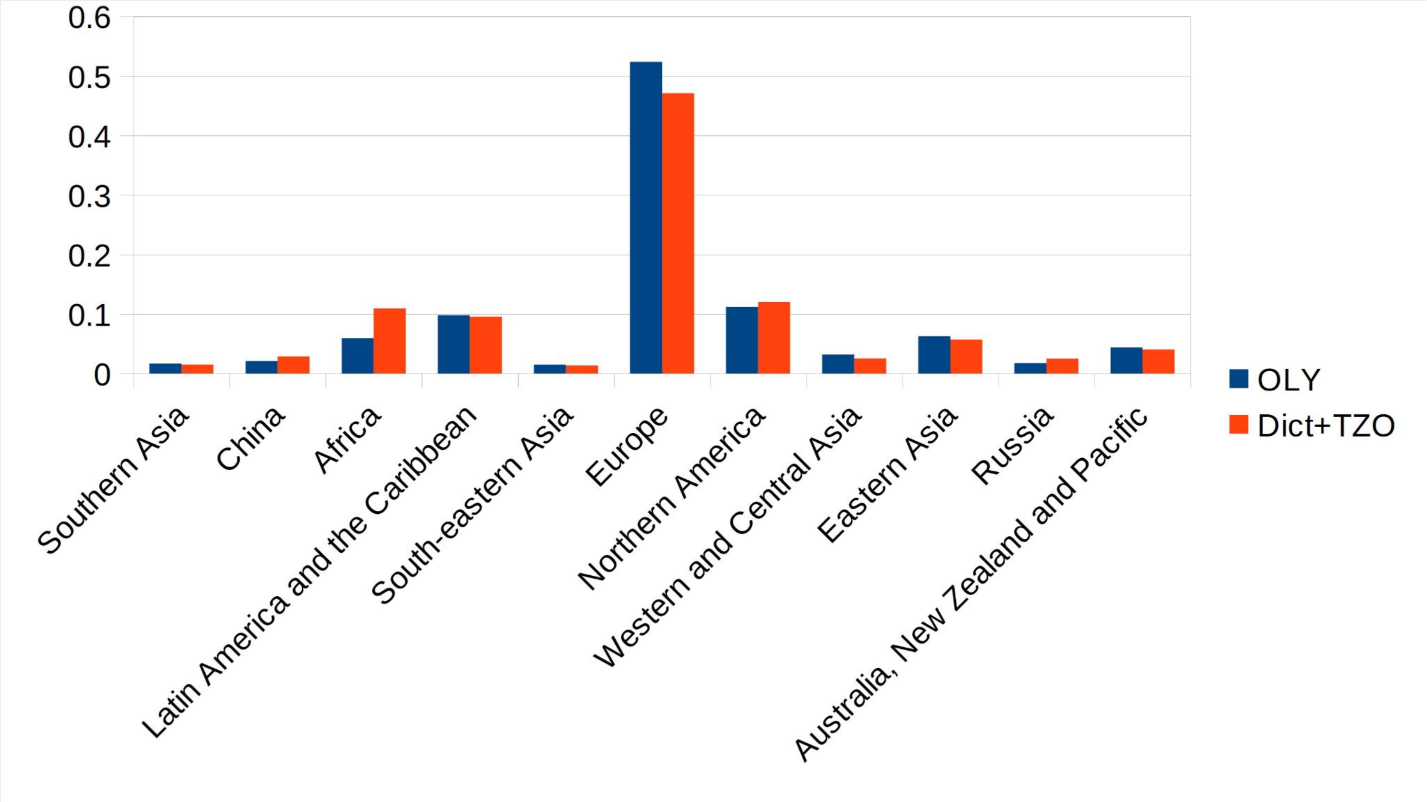
Olympics: TZO offset only



Olympics: dictionary + TZO

	precision	recall	f1-score	support
Europe	0.984	0.821	0.896	66408
Northern America	0.891	0.884	0.887	14194
Latin America and the Caribbean	0.908	0.821	0.862	12410
Eastern Asia	0.996	0.841	0.912	7931
Africa	0.468	0.803	0.591	7485
Australia, New Zealand and Pacific	0.994	0.853	0.918	5529
Western and Central Asia	0.829	0.612	0.704	4013
China	0.783	0.991	0.875	2653
Russia	0.628	0.833	0.716	2203
Southern Asia	0.987	0.835	0.905	2090
South-eastern Asia	0.730	0.609	0.664	1887
[UNK]	0.000	0.000	0.000	0
accuracy	0.824	0.824	0.824	0
macro avg	0.766	0.742	0.744	126803
weighted avg	0.918	0.824	0.863	126803

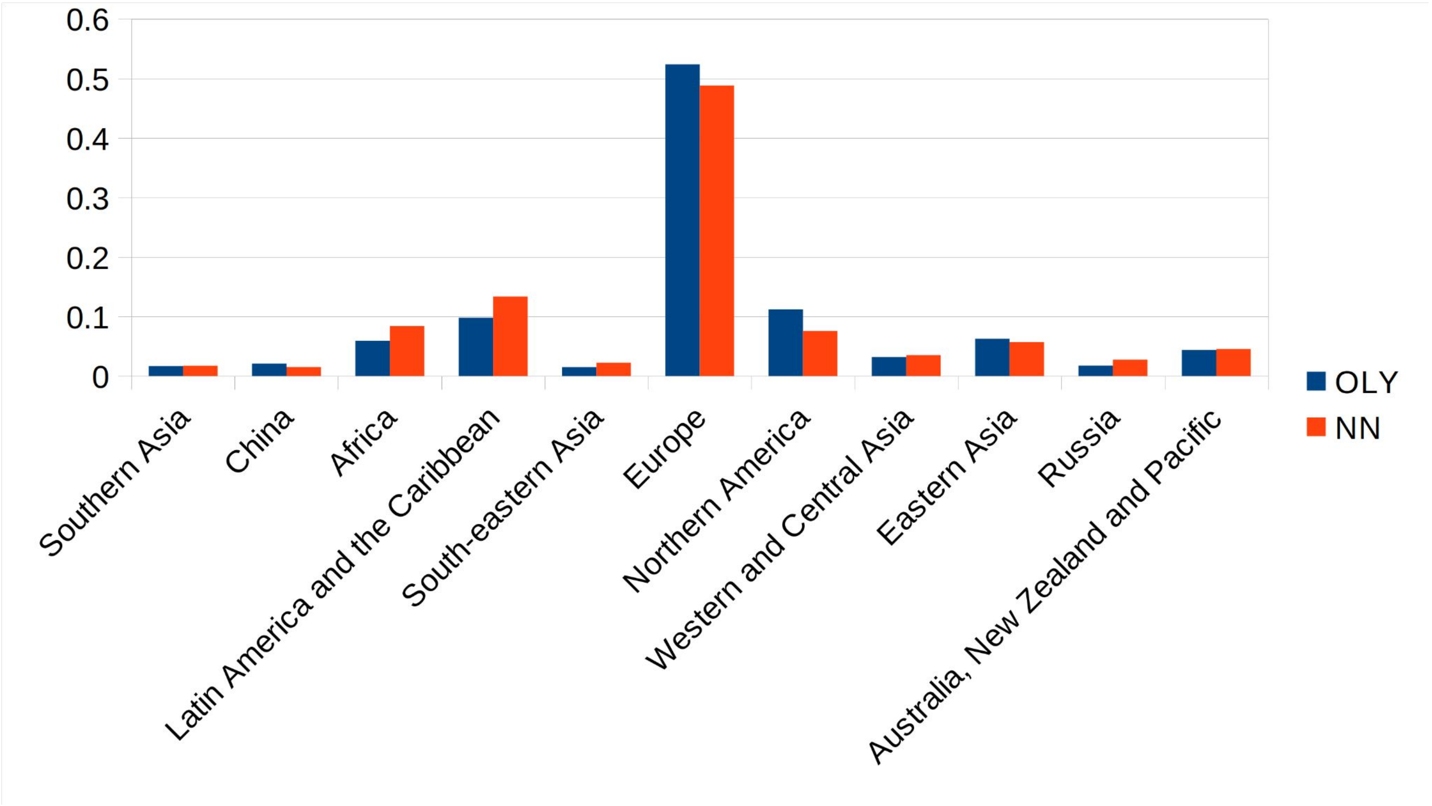
Olympics: dictionary + TZO



Olympics: NN (multiple TZO models)

	precision	recall	f1-score	support
Europe	0.962	0.897	0.928	66408
Northern America	0.894	0.604	0.721	14194
Latin America and the Caribbean	0.668	0.911	0.771	12410
Eastern Asia	0.884	0.804	0.842	7931
Africa	0.515	0.732	0.604	7485
Australia, New Zealand and Pacific	0.919	0.955	0.936	5529
Western and Central Asia	0.538	0.595	0.565	4013
China	0.713	0.510	0.595	2653
Russia	0.398	0.628	0.487	2203
Southern Asia	0.955	0.991	0.973	2090
South-eastern Asia	0.584	0.871	0.699	1887
[UNK]	0.000	0.000	0.000	0
accuracy	0.831	0.831	0.831	0
macro avg	0.669	0.708	0.677	126803
weighted avg	0.858	0.831	0.837	126803

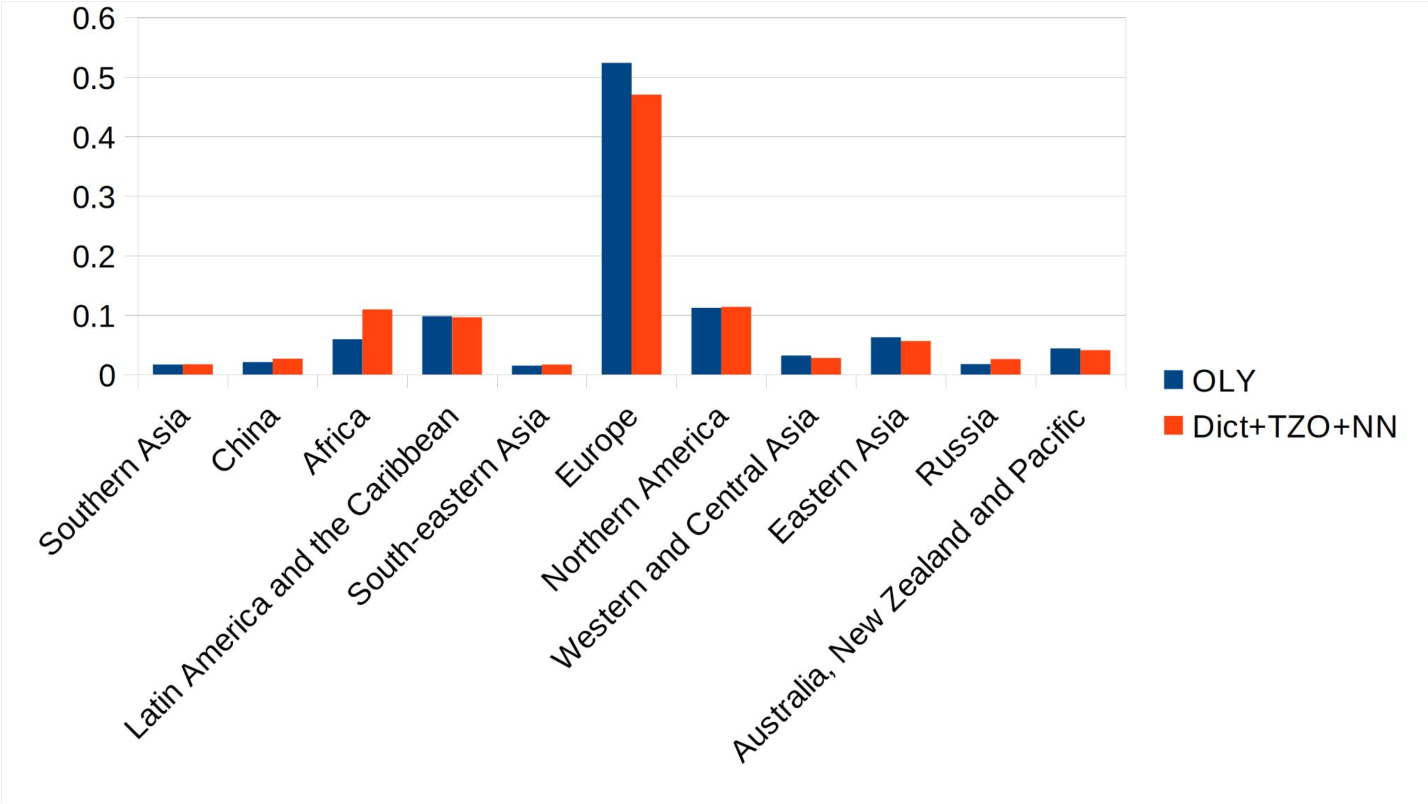
Olympics: NN (multiple TZO models)



Olympics: dictionary + TZO + NN

	precision	recall	f1-score	support
Europe	0.980	0.880	0.927	66408
Northern America	0.888	0.902	0.895	14194
Latin America and the Caribbean	0.886	0.871	0.878	12410
Eastern Asia	0.995	0.894	0.942	7931
Africa	0.465	0.860	0.604	7485
Australia, New Zealand and Pacific	0.988	0.923	0.954	5529
Western and Central Asia	0.767	0.674	0.717	4013
China	0.783	0.991	0.875	2653
Russia	0.584	0.869	0.699	2203
Southern Asia	0.969	0.995	0.982	2090
South-eastern Asia	0.780	0.870	0.823	1887
[UNK]	0.000	0.000	0.000	0
accuracy	0.880	0.880	0.880	0
macro avg	0.757	0.811	0.775	126803
weighted avg	0.911	0.880	0.890	126803

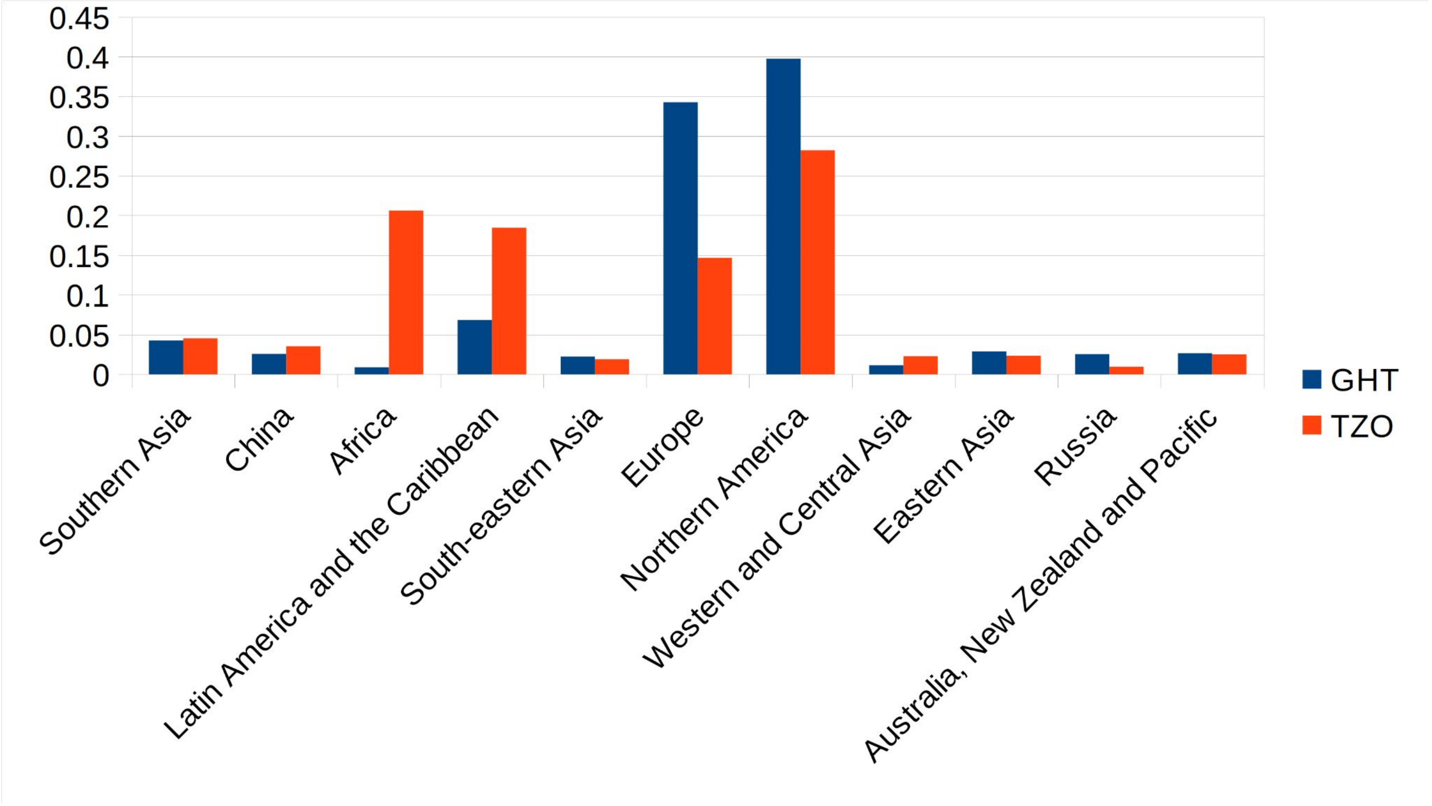
Olympics: dictionary + TZO + NN



GHT: TZO offset only

	precision	recall	f1-score	support
Northern America	0.965	0.684	0.800	96273
Europe	0.961	0.411	0.575	82909
Latin America and the Caribbean	0.318	0.855	0.463	16572
Southern Asia	0.921	0.981	0.950	10272
Eastern Asia	0.962	0.786	0.865	6936
Australia, New Zealand and Pacific	0.977	0.923	0.949	6439
China	0.614	0.855	0.715	6117
Russia	0.359	0.140	0.202	5958
South-eastern Asia	0.718	0.610	0.660	5390
Western and Central Asia	0.164	0.328	0.218	2749
Africa	0.024	0.575	0.046	2099
[UNK]	0.000	0.000	0.000	0
accuracy	0.608	0.608	0.608	0
macro avg	0.582	0.596	0.537	241714
weighted avg	0.871	0.608	0.679	241714

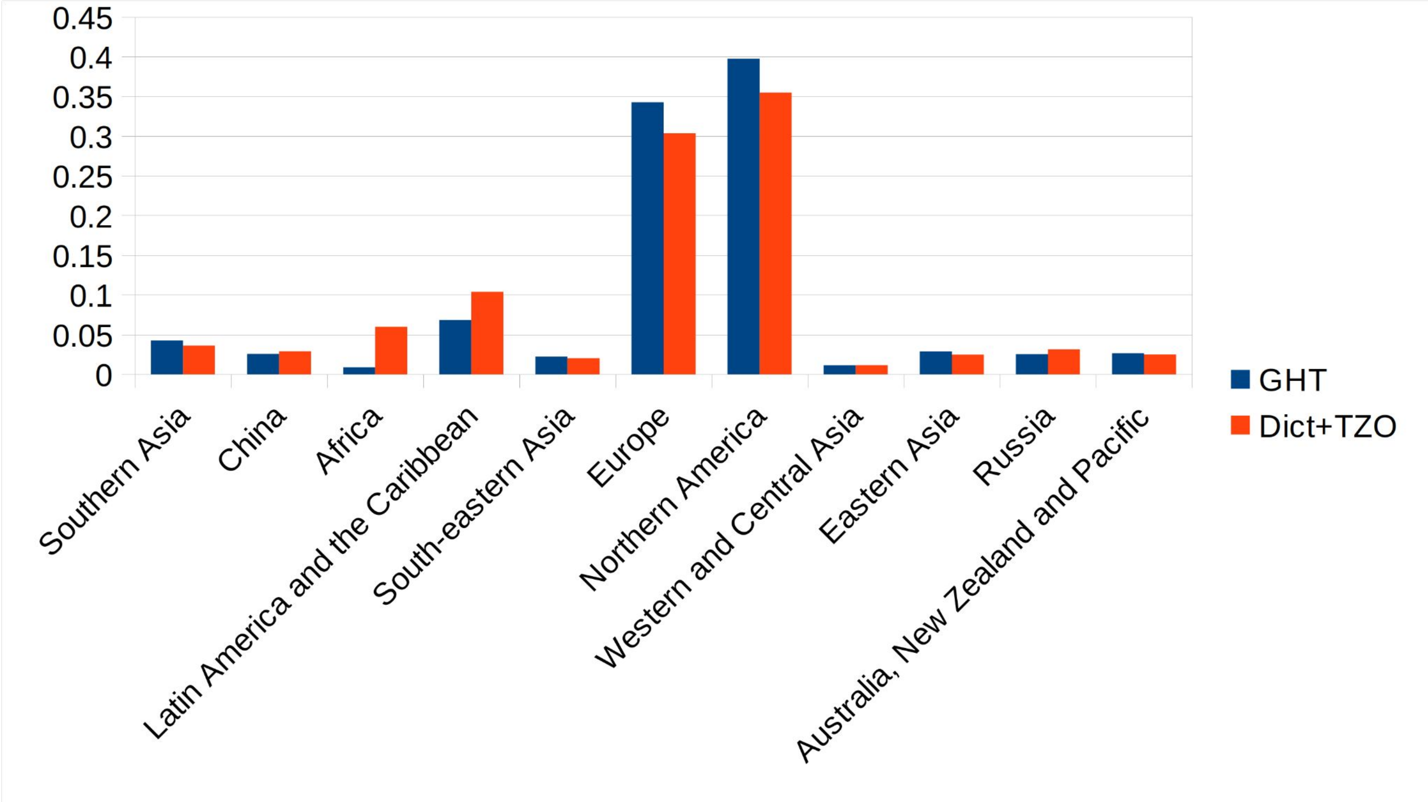
GHT: TZO offset only



GHT: dictionary + TZO

	precision	recall	f1-score	support
Northern America	0.990	0.785	0.876	96273
Europe	0.990	0.779	0.872	82909
Latin America and the Caribbean	0.664	0.897	0.763	16572
Southern Asia	0.999	0.756	0.861	10272
Eastern Asia	0.942	0.725	0.819	6936
Australia, New Zealand and Pacific	0.949	0.797	0.867	6439
China	0.786	0.802	0.794	6117
Russia	0.666	0.754	0.707	5958
South-eastern Asia	0.771	0.619	0.686	5390
Western and Central Asia	0.710	0.640	0.673	2749
Africa	0.116	0.712	0.200	2099
[UNK]	0.000	0.000	0.000	0
accuracy	0.782	0.782	0.782	0
macro avg	0.715	0.689	0.676	241714
weighted avg	0.937	0.782	0.846	241714

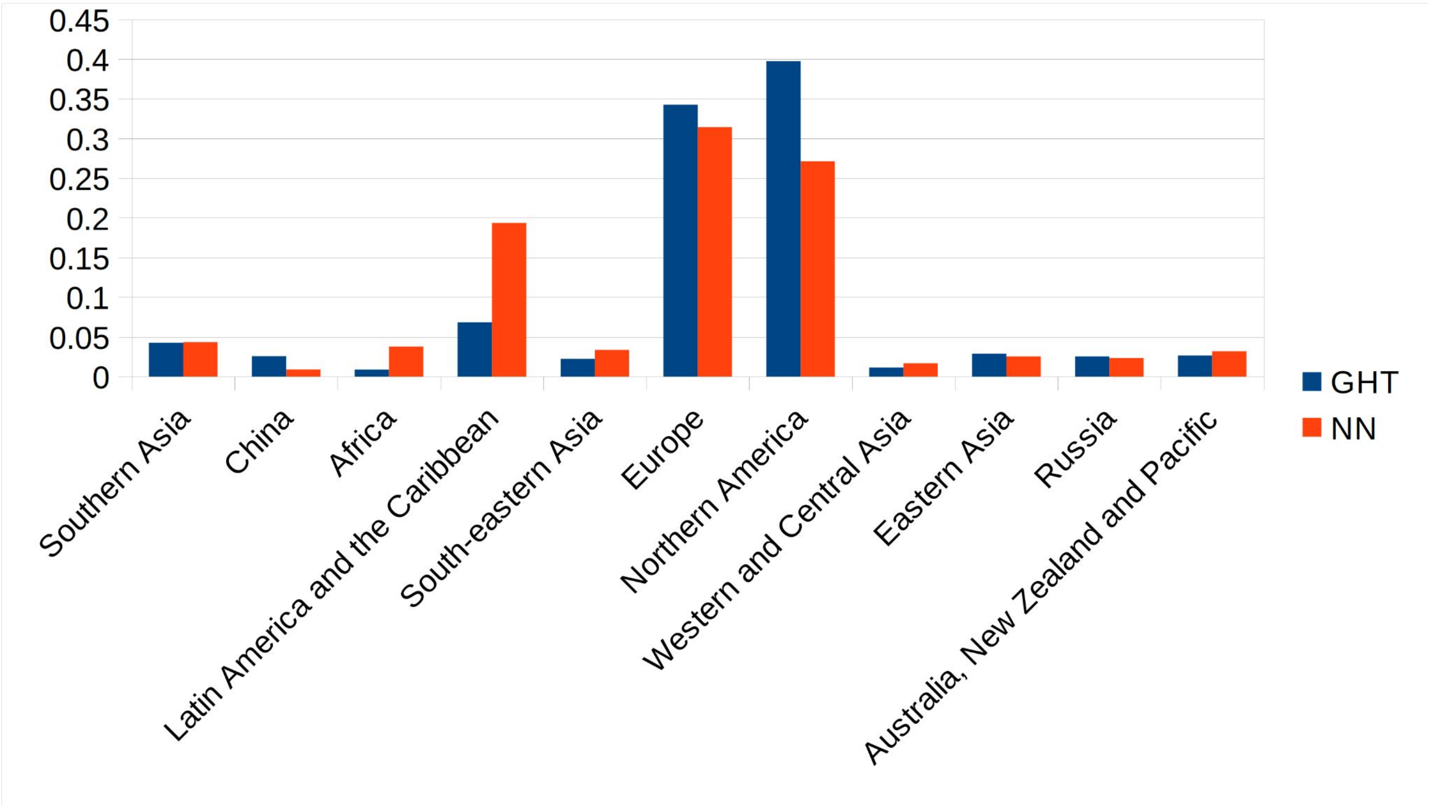
GHT: dictionary + TZO



GHT: NN (multiple TZO models)

	precision	recall	f1-score	support
Northern America	0.987	0.672	0.800	96273
Europe	0.962	0.882	0.920	82909
Latin America and the Caribbean	0.326	0.920	0.482	16572
Southern Asia	0.977	0.997	0.987	10272
Eastern Asia	0.779	0.690	0.732	6936
Australia, New Zealand and Pacific	0.764	0.913	0.831	6439
China	0.740	0.259	0.384	6117
Russia	0.611	0.582	0.596	5958
South-eastern Asia	0.548	0.823	0.658	5390
Western and Central Asia	0.417	0.612	0.496	2749
Africa	0.127	0.552	0.207	2099
accuracy	0.771	0.771	0.771	0
macro avg	0.658	0.718	0.645	241714
weighted avg	0.881	0.771	0.799	241714

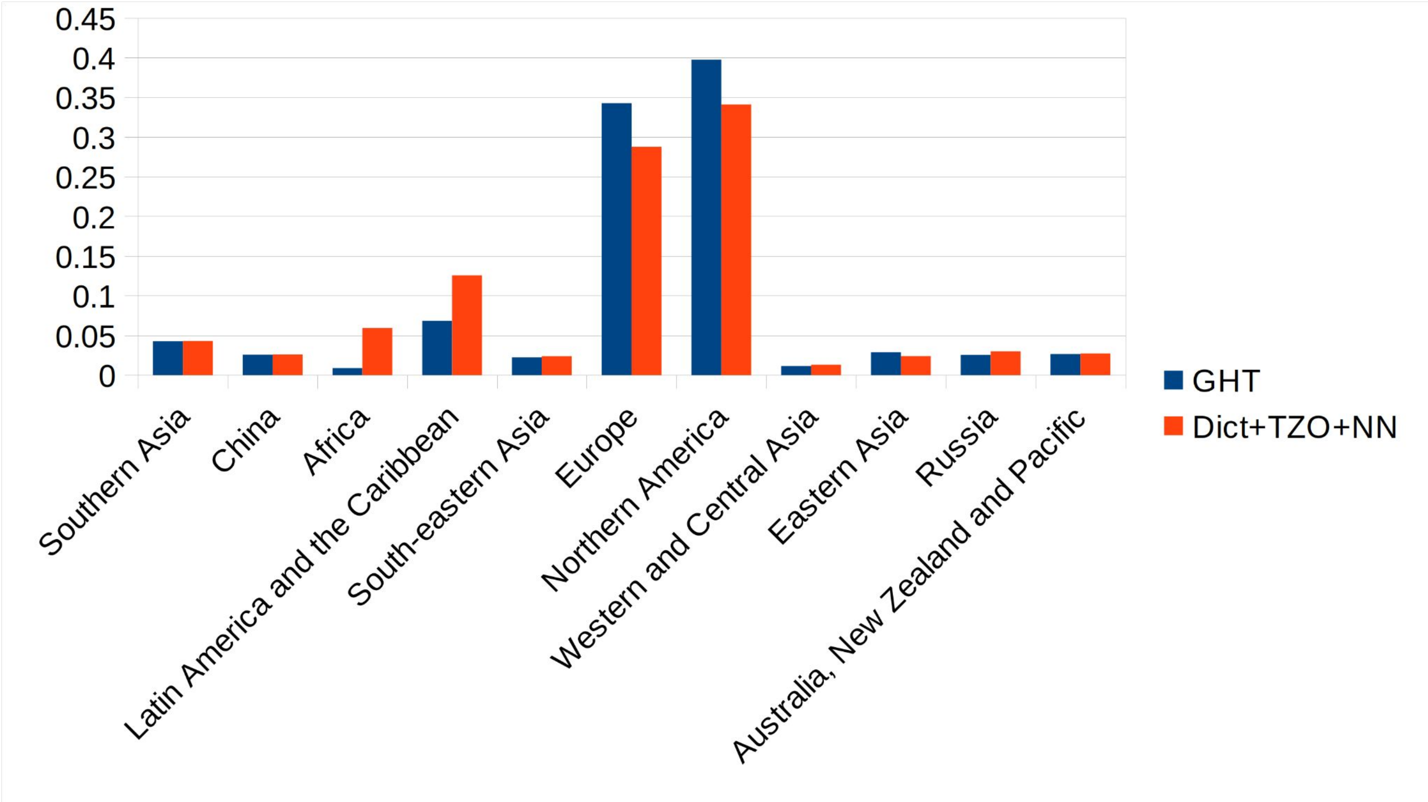
GHT: NN (multiple TZO models)



GHT: dictionary + TZO + NN

	precision	recall	f1-score	support
Northern America	0.989	0.847	0.912	96273
Europe	0.985	0.827	0.899	82909
Latin America and the Caribbean	0.515	0.943	0.666	16572
Southern Asia	0.990	0.998	0.994	10272
Eastern Asia	0.933	0.774	0.846	6936
Australia, New Zealand and Pacific	0.906	0.928	0.917	6439
China	0.784	0.803	0.794	6117
Russia	0.637	0.769	0.697	5958
South-eastern Asia	0.772	0.821	0.796	5390
Western and Central Asia	0.628	0.716	0.669	2749
Africa	0.113	0.770	0.197	2099
accuracy	0.847	0.847	0.847	0
macro avg	0.750	0.836	0.762	241714
weighted avg	0.921	0.847	0.873	241714

GHT: dictionary + TZO + NN



Conclusions

It is feasible.

Current work:

- shrink the zones;
- improve the classifiers;
- mix the methods in a smarter way.

Thanks for listening