UNIVERSITÀ
DEGLI STUDI
DI PADOVA

The importance of being earnest (and average)

Arjuna Tuzzi
arjuna.tuzzi@unipd.it

Similarity Measures

A measure of similarity should mirror the proximity of two texts.

This numeric value depends on the linguistic features of texts and **on the measure** itself.

A pairwise measure of similarity should mirror the proximity between all texts (all pairs of texts) included in a corpus.

Choosing an appropriate (dis)similarity measure is crucial in many application domains, e.g. **text clustering** and **authorship attribution**.

Hundreds of different measures are available (cfr. Rudman 1998; Stamatatos 2009) but no measure can be considered best suited for all applications.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions, *Computers and the Humanities*, 31: 351-365.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for Information Science and Technology*, 60(3): 538-556.

many distances...

Cosine similarity:

$$c(A,B)=\frac{\vec{v}_A\cdot\vec{v}_B}{\left|\vec{v}_A\right|\times\left|\vec{v}_B\right|}$$

Delta distance:

$$\Delta(A,B)=\frac{1}{m}\sum_{i=1}^{m}\left|z_{iA}-z_{iB}\right| \qquad z_{ij}=\frac{f_{ij}-\mu_i}{\sigma_i}$$

Labbé's distance:

$$d(A,B)=\frac{\sum\limits_{i\in V_{A\cup B}}\left|f_{i,A}-f^{*}_{i,B}\right|}{2N_A} \qquad f^{*}_{i,B}=f_{i,B}\,N_A/N_B$$

(...)

Corpus of Italian Contemporary Novels

Our corpus includes **150 novels** (nearly 10 millions word-tokens)
written by **40 different authors**:

Affinati, Ammaniti, Bajani, Balzano, Baricco, Benni, Brizzi, Carofiglio, Covacich, De Luca, De Silva, Faletti, Ferrante, Fois, Giordano, Lagioia, Maraini, Mazzantini, Mazzucco, Milone, Montesano, Morazzoni, Murgia, Nesi, Nori, Parrella, Piccolo, Pincio, Prisco, Raimo, Ramondino, Rea, Scarpa, Sereni, Starnone, Tamaro, Valerio, Vasta, Veronesi, Vinci.

Language: all novels were originally written in Italian

Time: all novels have been published in the time span [1987-2016]
exceptions: Prisco 1966, *Una spirale di nebbia*; Prisco 1969, *La provincia addormentata*; Maraini 1972, *Memorie di una ladra*; Morazzoni 1986, *La ragazza col turbante*

Target: all authors are novelists and their works were written for adult readers

A. Tuzzi, M.A. Cortelazzo (2018), What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first 19 January 2018 fqx066, https://doi.org/10.1093/llc/fqx066).

A. Tuzzi, M.A. Cortelazzo (2018, eds), *Proceedings of the Workshop Drawing Elena Ferrante's Profile Padova, 7 September 2017*, Padova University Press, Padova (ISBN: 978-88-6938-130-0). http://www.padovauniversitypress.it/publications/9788869381300 (**free download**)

Distance

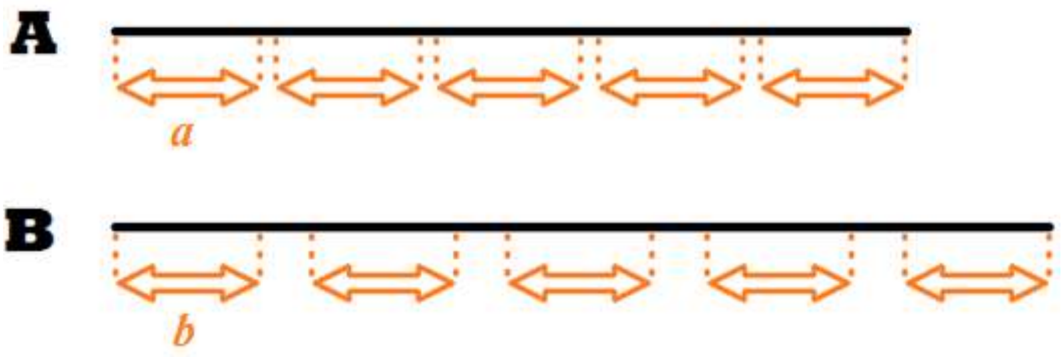
Labbé's intertextual distance:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^{*}_{i,B} \right|}{2N_A}$$

$$f^{*}_{i,B} = f_{i,B}\, N_A / N_B$$

Iterative version of distance based on equal-sized chunks

$$d_j(a \in A, b \in B) = \frac{\sum_{i \in V_{a \cup b}} \left| f_{i,a} - f_{i,b} \right|}{2n}$$



$$\hat{d}(A,B) = \frac{\sum_{j=1}^{k} d_j}{k}$$

n is the size in word tokens of chunks
k is the number of replications

Distance

Corpus of p = 150 novels:

- k = **500 replications**
- n = **10,000 word tokens**
- taking into account all word types of the vocabulary (149,870 items)

we obtained a square matrix that includes 150 x 150 cells
and 11,175 non-zero non-redundant values [$p(p-1)/2$].

When you have a distance between pairs of novels:

1. you have the base for a cluster analysis
2. you have a **ranking system**

M.A. Cortelazzo, P. Nadalutti, A. Tuzzi (2013), Improving Labbé’s Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature, *Journal of Quantitative Linguistics*, 20(2), pp. 125-152.

A. Tuzzi (2010), What to put in the bag? Comparing and contrasting procedures for text clustering, *Italian Journal of Applied Statistics / Statistica Applicata*, 22(1), pp. 77-94

Distance and Ranking

	Novel 1	Novel 2	Novel 3	Novel 4	Novel 5	Novel 6	Novel 7	Novel 8	Novel 9	Novel10	…
Novel 1	0	0.4760	0.5545	0.5545	0.5786	0.5505	0.5533	0.5533	0.5661	0.5373	…
Novel 2	0.4760	0	0.5270	0.5202	0.5559	0.5191	0.5304	0.5420	0.5349	0.5095	…
Novel 3	0.5545	0.5270	0	0.4166	0.4552	0.4147	0.4790	0.4783	0.5037	0.4735	…
Novel 4	0.5545	0.5202	0.4167	0	0.4381	0.3965	0.4731	0.4681	0.4888	0.4583	…
Novel 5	0.5786	0.5559	0.4552	0.4381	0	0.4455	0.4045	0.4370	0.4755	0.4989	…
Novel 6	0.5505	0.5191	0.4147	0.3965	0.4456	0	0.4708	0.4601	0.4826	0.4618	…
Novel 7	0.5533	0.5304	0.4790	0.4731	0.4045	0.4708	0	0.3653	0.4087	0.4807	…
Novel 8	0.5533	0.5420	0.4783	0.4681	0.4370	0.4601	0.3653	0	0.4198	0.4790	…
Novel 9	0.5661	0.5349	0.5037	0.4888	0.4755	0.4826	0.4087	0.4198	0	0.4990	…
Novel10	0.5373	0.5095	0.4735	0.4583	0.4989	0.4618	0.4807	0.4790	0.4990	0	…
…	…	…	…	…	…	…	…	…	…	…	…

Rows (or columns) represent a **ranking system**,
i.e. I can choose a novel and sort all the others from the closest to the furthest.

Ranking: Ermanno Rea

Mistero napoletano	*(Rea 1995)*	*La dismissione*	*(Rea 2002)*	*La comunista*	*(Rea 2012)*
Rea 1995	0	Rea 2002	0	Rea 2012	0
Rea 2002	0.418	Rea 2012	0.398	Rea 2002	0.398
Rea 2012	0.420	Tamaro 1994	0.415	Rea 1995	0.420
Pincio 2011	0.441	Murgia 2015	0.417	Tamaro 2006	0.431
Giordano 2014	0.442	Rea 1995	0.418	Tamaro 1994	0.431
Piccolo 2013	0.445	Tamaro 2006	0.426	Tamaro 1994	0.434
Tamaro 1994	0.446	Giordano 2014	0.429	Morazzoni 2005	0.437
Tamaro 2006	0.447	Carofiglio 2013	0.430	Giordano 2014	0.439
Tamaro 1994	0.447	Carofiglio 2004	0.430	Pincio 2012	0.442
Pincio 2012	0.451	Carofiglio 2003	0.432	Carofiglio 2013	0.444
Morazzoni 2005	0.451	Tamaro 2013	0.433	Starnone 2007	0.444
Tamaro 2013	0.453	Carofiglio 2010	0.433	Tamaro 2013	0.445
Carofiglio 2013	0.454	Carofiglio 2011	0.434	Carofiglio 2004	0.446
Carofiglio 2014	0.457	Ferrante 2014	0.436	Pincio 2011	0.447
Starnone 2007	0.458	Sereni 2015	0.436	Carofiglio 2011	0.448
De Silva 2011	0.459	Ferrante 2013	0.438	Ferrante 2014	0.449
Carofiglio 2006	0.460	Carofiglio 2014	0.438	Faletti 2006	0.450
Carofiglio 2003	0.460	Carofiglio 2006	0.438	Starnone 2011	0.451
Carofiglio 2011	0.460	Pincio 2012	0.438	Faletti 2009	0.452
Veronesi 1995	0.460	Starnone 2007	0.438	Ferrante 2012	0.452

Ranking: Ermanno Rea

Mistero napoletano	*(Rea 1995)*	
Rea 1995	**0**	**1st**
Rea 2002	**0.418**	**2nd**
Rea 2012	**0.420**	**3rd**
Pincio 2011	0.441	
Giordano 2014	0.442	
Piccolo 2013	0.445	
Tamaro 1994	0.446	
Tamaro 2006	0.447	
Murgia 2015	0.447	
Pincio 2012	0.451	
Morazzoni 2005	0.451	
Tamaro 2013	0.453	
Carofiglio 2013	0.454	
Carofiglio 2014	0.457	
Starnone 2007	0.458	
De Silva 2011	0.459	
Carofiglio 2006	0.460	
Carofiglio 2003	0.460	
Carofiglio 2011	0.460	
Veronesi 1995	0.460	

La dismissione	*(Rea 2002)*	
Rea 2002	**0**	**1st**
Rea 2012	**0.398**	**2nd**
Tamaro 1994	0.415	
Murgia 2015	0.417	
Rea 1995	**0.418**	**5th**
Tamaro 2006	0.426	
Giordano 2014	0.429	
Carofiglio 2013	0.430	
Carofiglio 2004	0.430	
Carofiglio 2003	0.432	
Tamaro 2013	0.433	
Carofiglio 2010	0.433	
Carofiglio 2011	0.434	
Ferrante 2014	0.436	
Sereni 2015	0.436	
Ferrante 2013	0.438	
Carofiglio 2014	0.438	
Carofiglio 2006	0.438	
Pincio 2012	0.438	
Starnone 2007	0.438	

La comunista	*(Rea 2012)*	
Rea 2012	**0**	**1st**
Rea 2002	**0.398**	**2nd**
Rea 1995	**0.420**	**3rd**
Tamaro 2006	0.431	
Tamaro 1994	0.431	
Murgia 2015	0.434	
Morazzoni 2005	0.437	
Giordano 2014	0.439	
Pincio 2012	0.442	
Carofiglio 2013	0.444	
Starnone 2007	0.444	
Tamaro 2013	0.445	
Carofiglio 2004	0.446	
Pincio 2011	0.447	
Carofiglio 2011	0.448	
Ferrante 2014	0.449	
Faletti 2006	0.450	
Starnone 2011	0.451	
Faletti 2009	0.452	
Ferrante 2012	0.452	

Evaluation of results

The square matrix reporting all pairwise distances may be interpreted according to its rows (or columns) as a ranking system: for each text, all other texts may be sorted from the closest to the furthest.

$p + 1 = 3$ novels are written by the same author (e.g. Ermanno Rea)

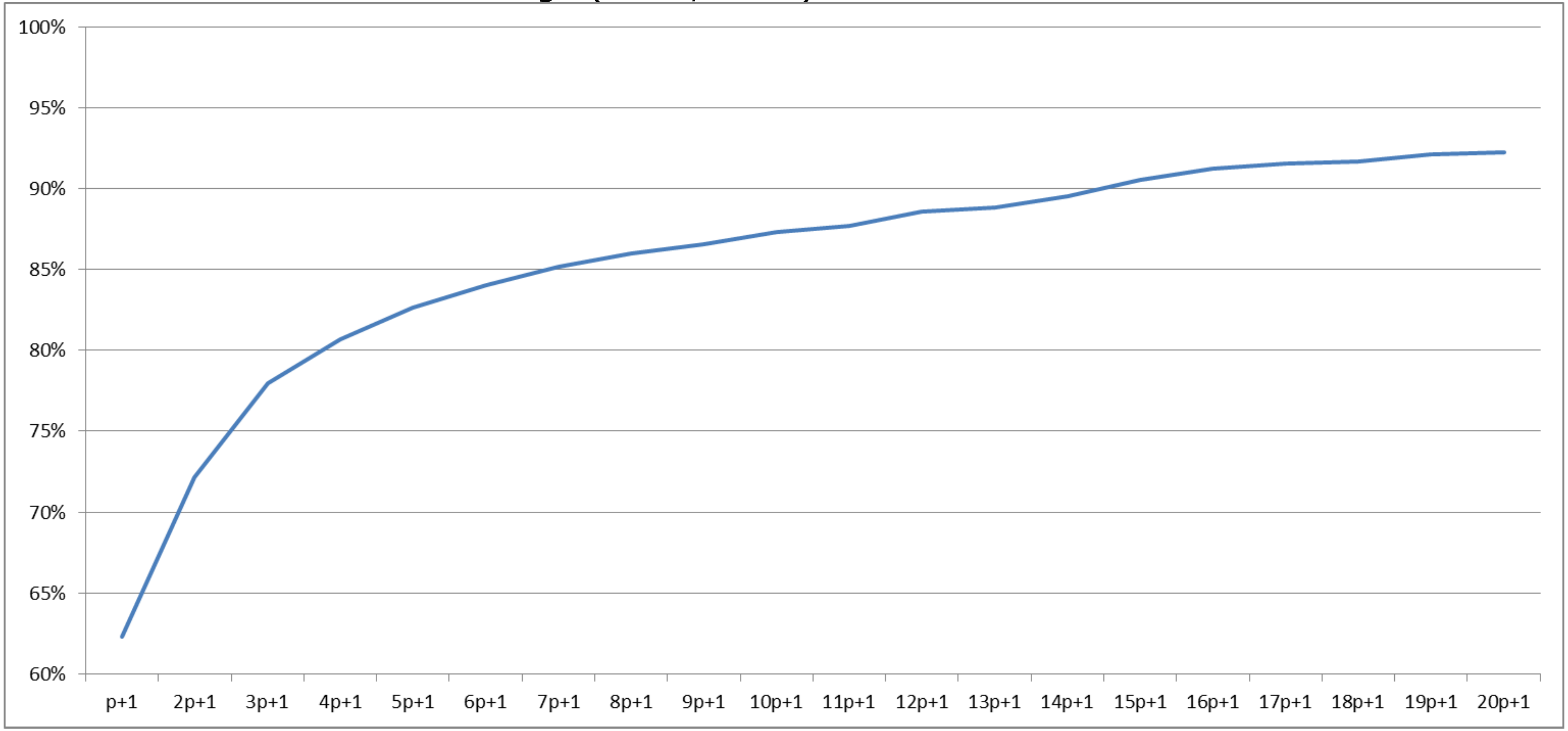
when we sort the 150 novels by increasing values of the distances from the first novel (e.g. *Mistero napoletano*, Rea 1995), all of Rea's novels are associated to ranks k between 1 and $p + 1$ (100% purity).

In Rea's case we have two novels that produce ranks between 1 and $p + 1 = 3$ and a novel that shows ranks between 1 and $2p + 1=5$

rank k	**Rea 1995**	**Rea 2002**	**Rea 2012**	**Tot**
$k \leq 3$	3	2	3	8
$4 \leq k \leq 5$	0	1	0	1
$k > 5$	0	0	0	0
	3	3	3	9
purity	100%	67%	100%	89%

Performance

Performance in terms of rankings (Tuzzi, 2010) for 150 novels



A. Tuzzi (2010), What to put in the bag? Comparing and contrasting procedures for text clustering, *Italian Journal of Applied Statistics / Statistica Applicata*, 22(1), pp. 77-94.

Ranks

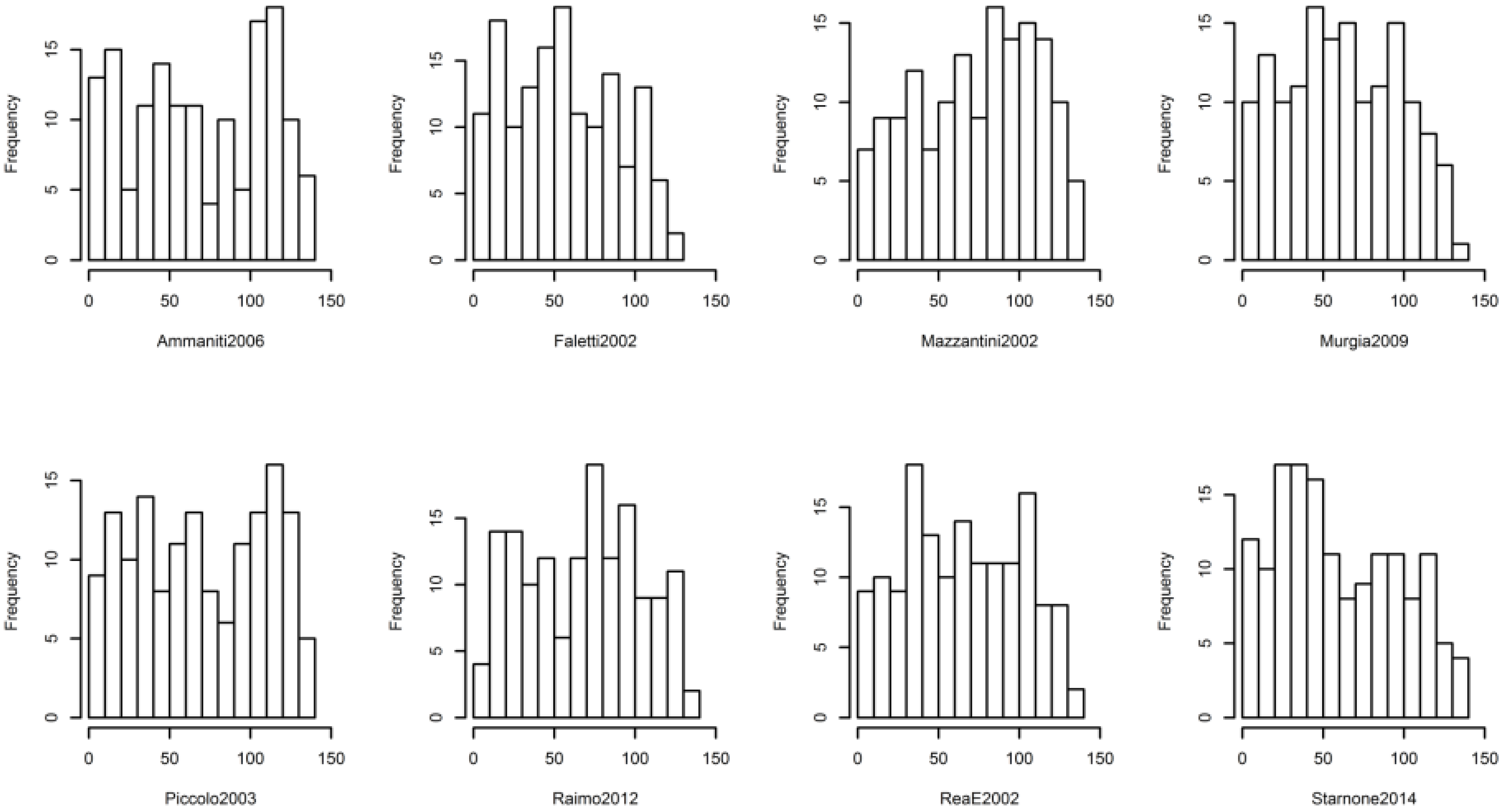
For each novel we have a ranking system (e.g. by column)

For each novel we can read positions in 150 rankings (e.g. by row)

	Affinati 1997	Affinati 2016	…	Rea 1995	Rea 2002	Rea 2012	…	Tamaro 1989	Tamaro 1991	Tamaro 1994	Tamaro 2006	Tamaro 2013	…
Affinati 1997	1	2	…	129	145	142	…	143	148	149	146	136	…
Affinati 2016	2	1	…	51	108	96	…	135	144	124	110	101	…
…	…	…	…				…						…
Rea 1995	20	11	…	1	5	3	…	94	130	81	68	44	…
Rea 2002	31	27	…	2	1	2	…	19	93	14	10	6	…
Rea 2012	39	26	…	3	2	1	…	50	121	39	16	21	…
…	…	…	…				…						…
Tamaro 1989	84	138	…	116	78	94	…	1	111	75	89	23	…
Tamaro 1991	113	111	…	83	51	78	…	16	1	3	4	13	…
Tamaro 1994	23	6	…	7	3	5	…	2	2	1	2	2	…
Tamaro 2006	7	3	…	8	6	4	…	5	12	2	1	3	…
Tamaro 2013	9	14	…	12	11	12	…	4	55	4	3	1	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…

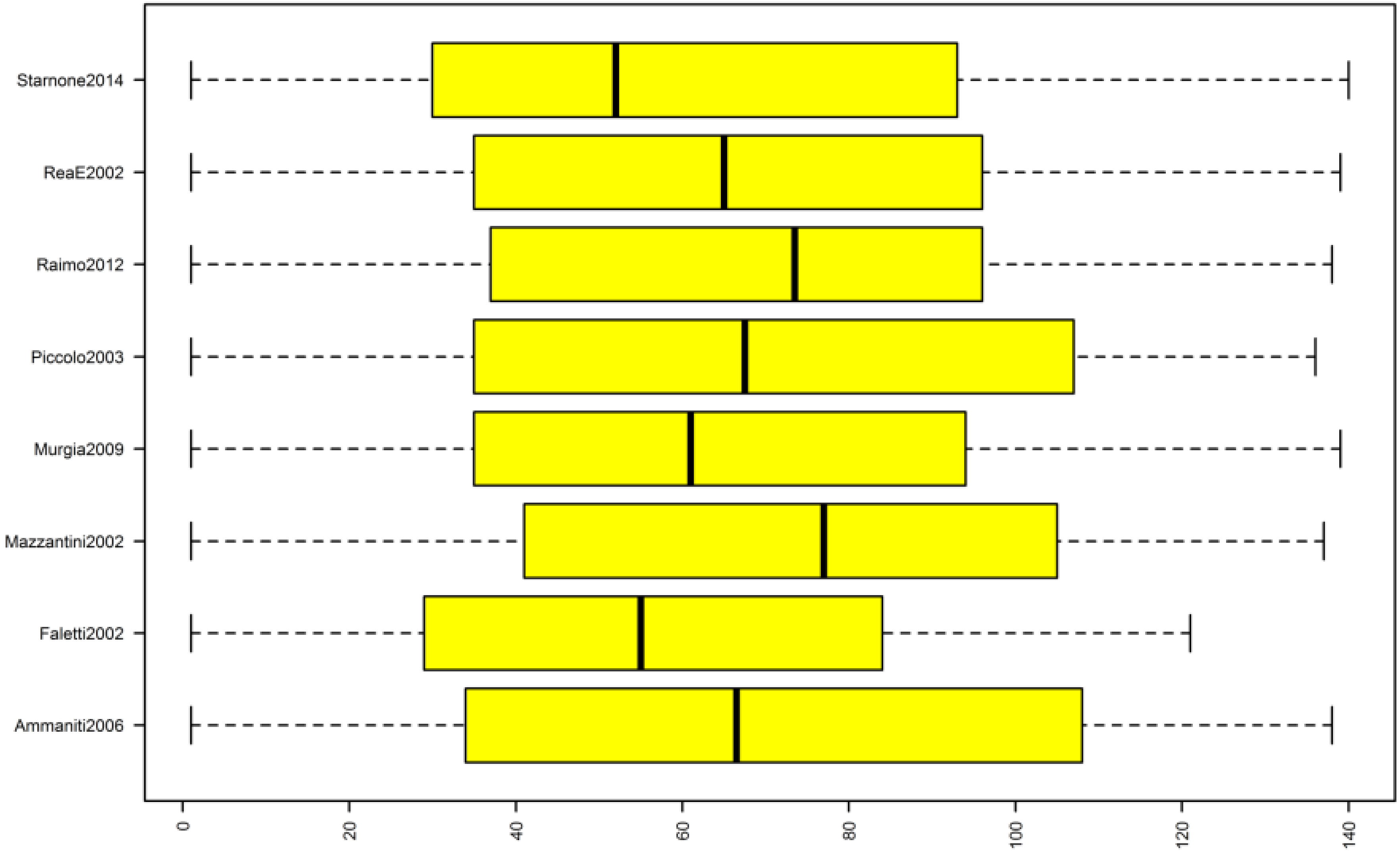
Examples

We expect to find a uniform distribution in [1..150]:

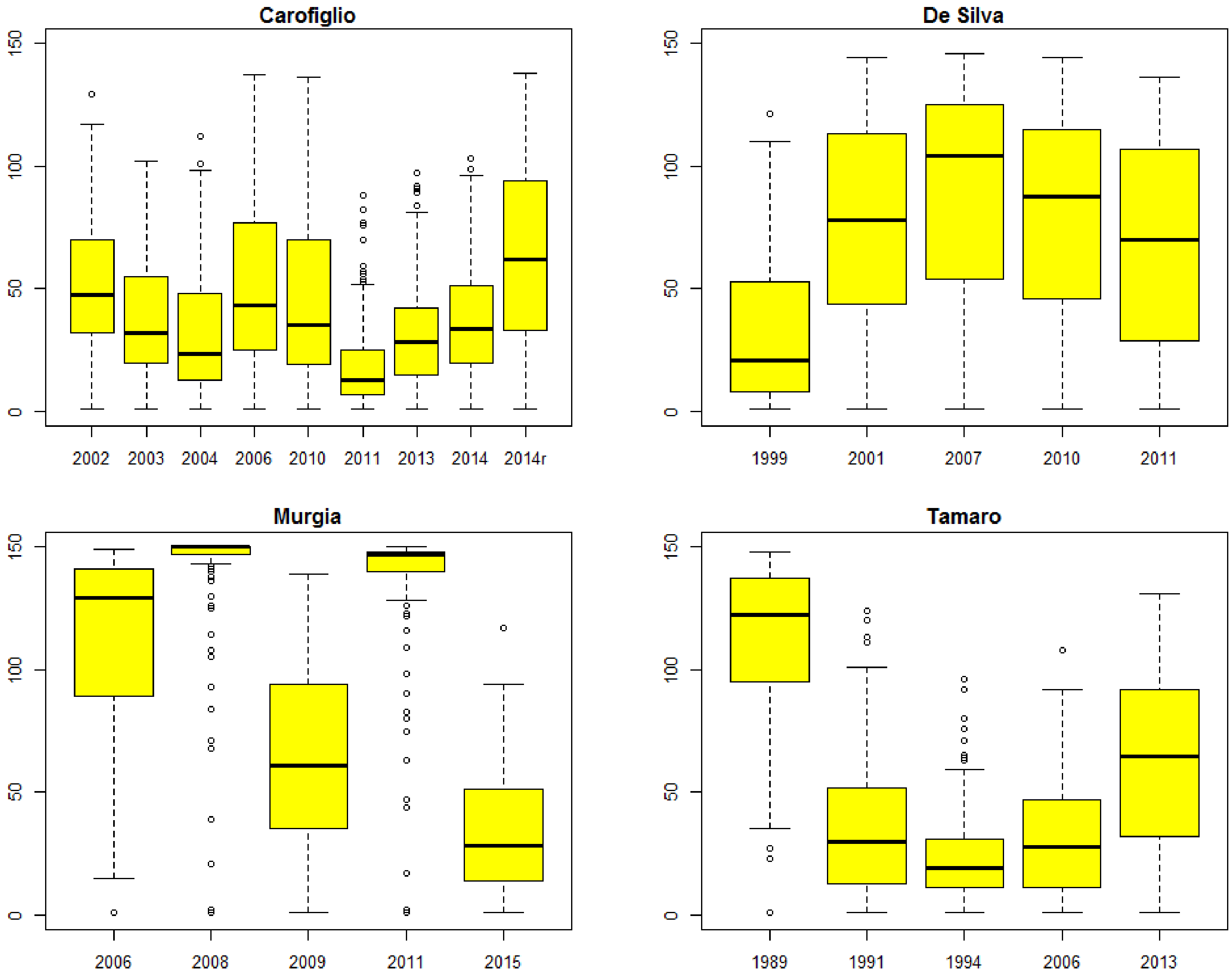


Examples

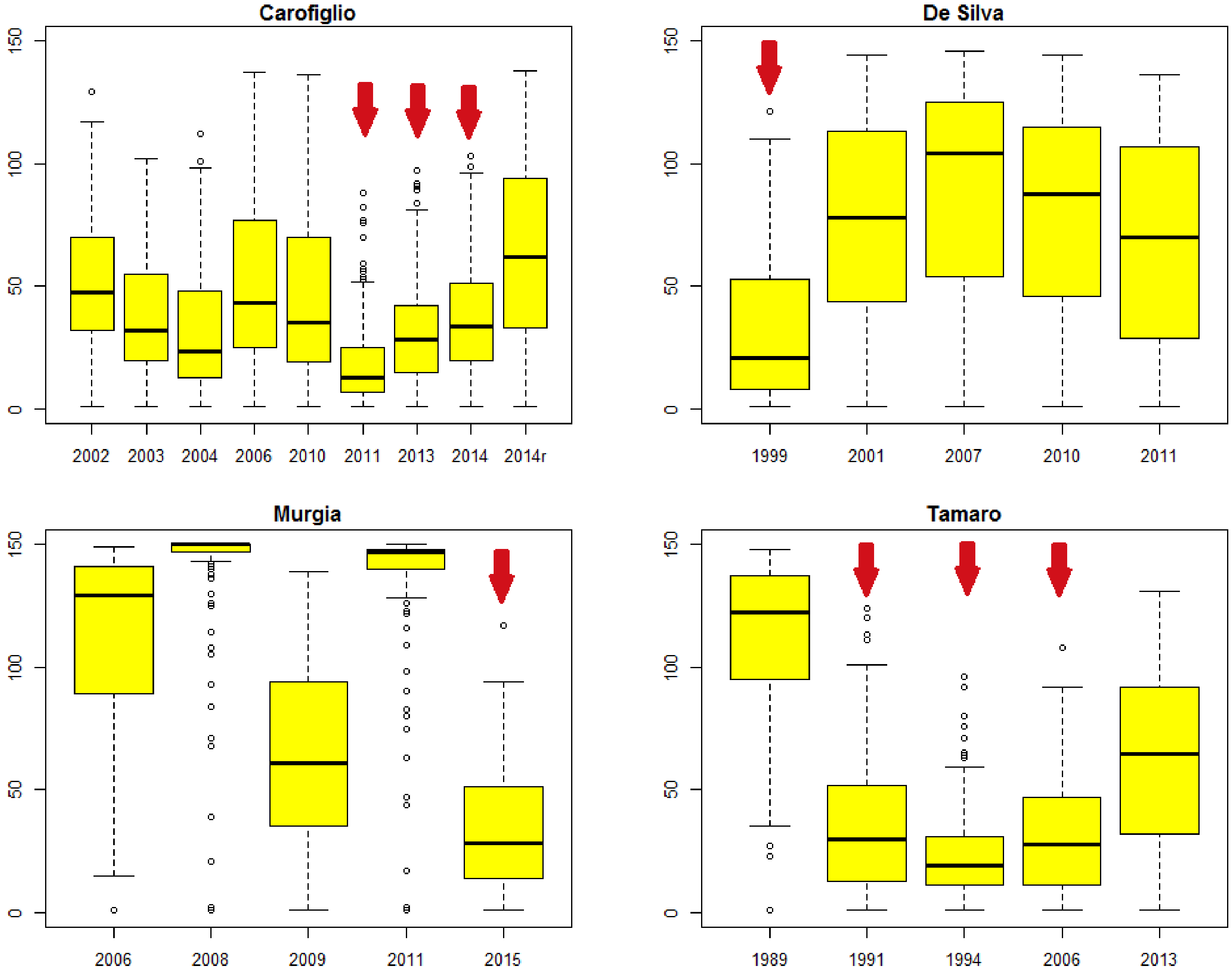
the same novels as box-plots:



"average" novels?



"average" novels?



Identifying "average" novels

Problem:

- we have a number of "average novels" that seem to be near to most of novels

How to identify them?

- median
- quartiles
- skewness
- distribution (we are testing Beta)

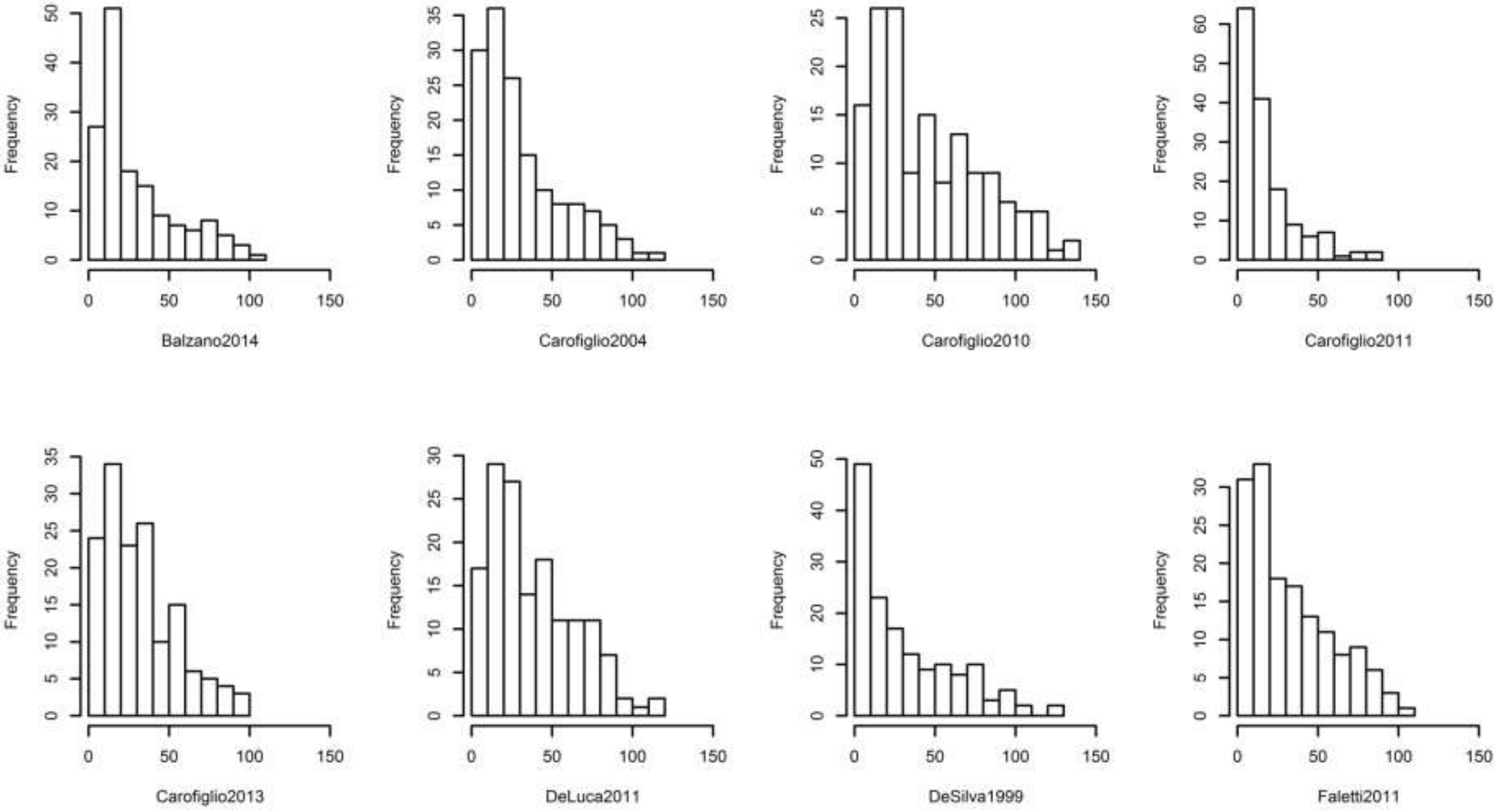
Idea:

- does the performance of a distance increase if we disregard these "average novels"?

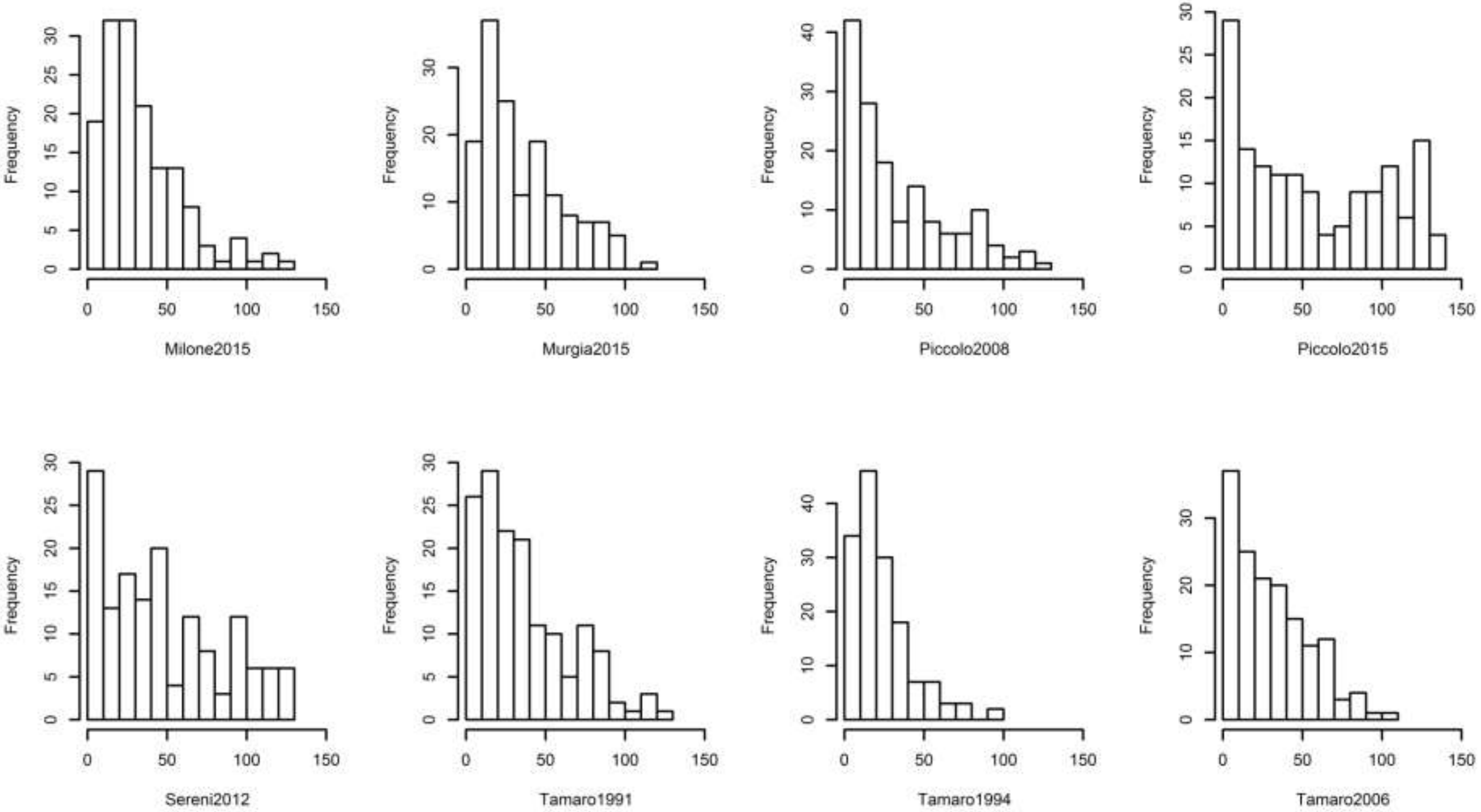
First example:

- we disregarded 16 (slightly more than 10%)

Identifying "average" novels



Identifying "average" novels



Effects on ranking: Ermanno Rea

Mistero napoletano	*(Rea 1995)*	
Rea 1995	**0**	1st
Rea 2002	**0.418**	2nd
Rea 2012	**0.420**	3rd
Pincio 2011	0.441	
Giordano 2014	0.442	
Piccolo 2013	0.445	
~~Tamaro 1994~~	~~0.446~~	
~~Tamaro 2006~~	~~0.447~~	
~~Murgia 2015~~	~~0.447~~	
Pincio 2012	0.451	
Morazzoni 2005	0.451	
Tamaro 2013	0.453	
~~Carofiglio 2013~~	~~0.454~~	
Carofiglio 2014	0.457	
Starnone 2007	0.458	
De Silva 2011	0.459	
Carofiglio 2006	0.460	
Carofiglio 2003	0.460	
~~Carofiglio 2011~~	~~0.460~~	
Veronesi 1995	0.460	

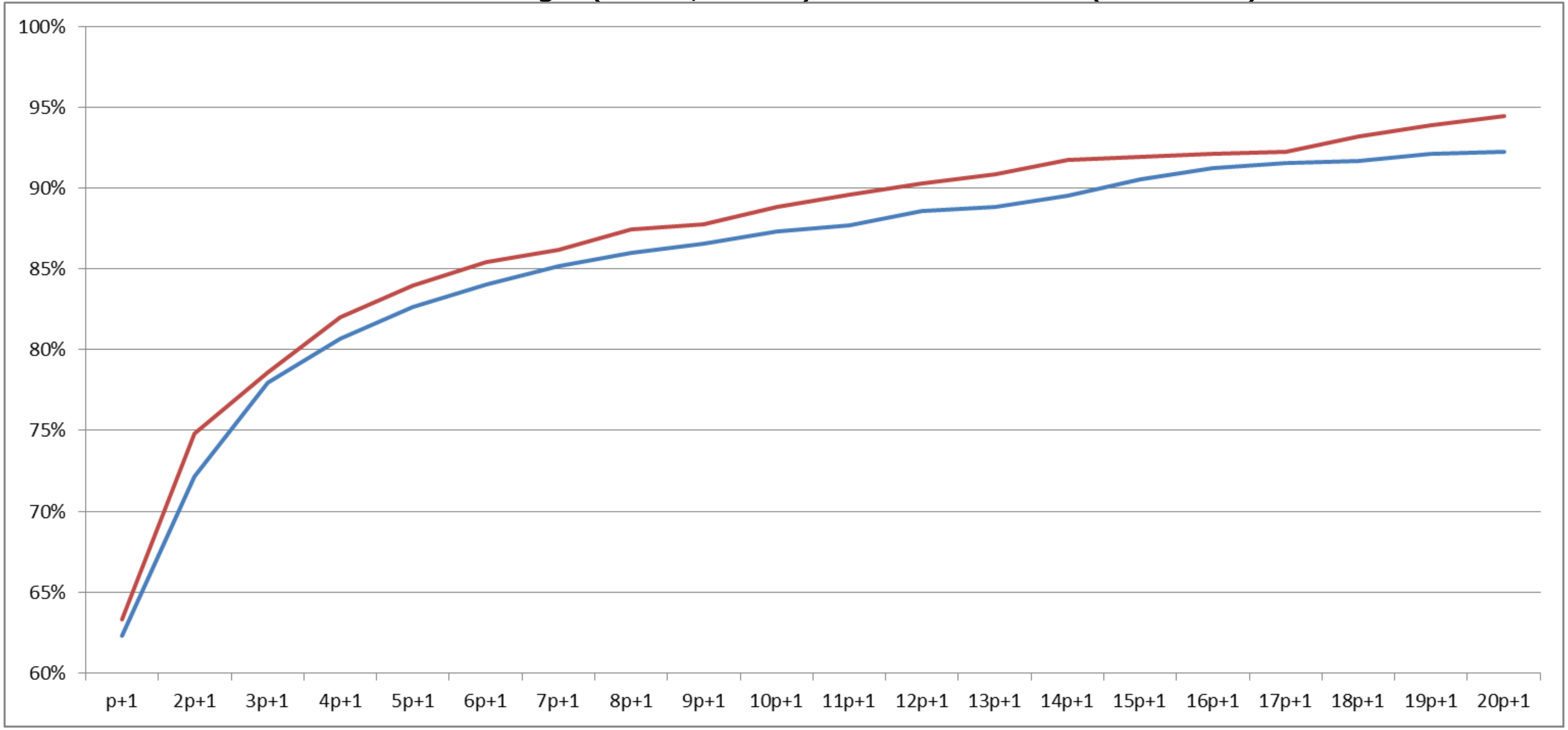
La dismissione	*(Rea 2002)*	
Rea 2002	**0**	1st
Rea 2012	**0.398**	2nd
~~Tamaro 1994~~	~~0.415~~	
~~Murgia 2015~~	~~0.417~~	
Rea 1995	**0.418**	3rd
~~Tamaro 2006~~	~~0.426~~	
Giordano 2014	0.429	
~~Carofiglio 2013~~	~~0.430~~	
~~Carofiglio 2004~~	~~0.430~~	
Carofiglio 2003	0.432	
Tamaro 2013	0.433	
~~Carofiglio 2010~~	~~0.433~~	
~~Carofiglio 2011~~	~~0.434~~	
Ferrante 2014	0.436	
Sereni 2015	0.436	
Ferrante 2013	0.438	
Carofiglio 2014	0.438	
Carofiglio 2006	0.438	
Pincio 2012	0.438	
Starnone 2007	0.438	

La comunista	*(Rea 2012)*	
Rea 2012	**0**	1st
Rea 2002	**0.398**	2nd
Rea 1995	**0.420**	3rd
~~Tamaro 2006~~	~~0.431~~	
~~Tamaro 1994~~	~~0.431~~	
~~Murgia 2015~~	~~0.434~~	
Morazzoni 2005	0.437	
Giordano 2014	0.439	
Pincio 2012	0.442	
~~Carofiglio 2013~~	~~0.444~~	
Starnone 2007	0.444	
Tamaro 2013	0.445	
~~Carofiglio 2004~~	~~0.446~~	
Pincio 2011	0.447	
~~Carofiglio 2011~~	~~0.448~~	
Ferrante 2014	0.449	
Faletti 2006	0.450	
Starnone 2011	0.451	
Faletti 2009	0.452	
Ferrante 2012	0.452	

Performance

We disregard 16 novels

Performance in terms of rankings (Tuzzi, 2010) for 134 novels (**red line**)



A. Tuzzi (2010), What to put in the bag? Comparing and contrasting procedures for text clustering, *Italian Journal of Applied Statistics / Statistica Applicata*, 22(1), pp. 77-94.

Concluding remarks

1. Are we able to identify the main distinctive feature of these "average novels"? Why are they close to all the others? (often they are best-sellers)

2. Are we able to identify the main distinctive features of "odd novels"? Why are they far from all the others?

3. Some distances produce this phenomenon and some distances do not.

4. What should we do when we have "average novels" among the candidates of an authorship attribution task (or part of a training set in a machine learning perspective)? If we disregard them, the performance of the distance improves.

Thank you!

www.giat.org