Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Large scale analysis of violent death count in daily newspapers to quantify bias and censorship

marco casolino (casolino@roma2.infn.it)

Research

Keywords: News, Death, Suicide, Scaling laws, Censorship, Gender bias, Newspaper

Posted Date: July 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17213/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published on August 11th, 2020. See the published version at https://doi.org/10.1186/s40537-020-00338-1.

RESEARCH

Large scale analysis of violent death count in daily newspapers to quantify bias and censorship.

Marco Casolino

Correspondence:
casolino@roma2.infn.it
INFN Structure of Roma Tor Vergata , Via della Ricerca Scientifica 1, Rome, Italy
Riken, Computational Astrophysics lab, Hirosawa, Wako, Japan
Dipartimento di Fisica, Università di Roma Tor Vergata, Via della Ricerca Scientifica 1, Rome, Italy
Full list of author information is available at the end of the article

Abstract

In this work we develop a series of techniques and tools to determine and quantify the presence of bias and censorship in newspapers. These algorithms are tested analyzing the occurrence of keywords 'killed' and 'suicide' (*'morti', 'suicidio'* in Italian) and their changes over time, gender and reported location on the complete online archives (42 million records) of the major US newspaper (*The New York Times*) and the three major Italian ones (*Il Corriere della Sera, La Repubblica, La Stampa*).

Using these tools, since the Italian language distinguishes between the female and male cases, we find the presence of gender bias in all Italian newspapers, with reported single female deaths to be about one-third of those involving single men.

Analyzing the historical trends, we show evidence of censorship in Italian newspapers both during World War 1 and during the Italian Fascist regime. Censorship in all countries during World Wars and in Italy during the Fascist period is a historically ascertained fact, but so far there was no estimate on the amount on censorship in newspaper reporting: in this work we estimate that about 75% of domestic deaths and suicides were not reported. This is also confirmed by statistical analysis of the distribution of the least significant digit of the number of reported deaths.

We also find that the distribution function of the number of articles vs. the number of deaths reported in articles follows a power law, which is broken (with fewer articles being written) when reporting on few deaths occurring in foreign countries. The lack of articles is found to grow with geographical distance from the nation where the newspaper is being printed.

Whereas the assessment of the truth of a single article or the debunking of what are now called 'fake news' requires specific fact-checking and becomes more difficult as time goes by, these methods can be be used in historical analysis and to evaluate quantitatively the amount of bias and censorship present in other printed or online publication and can thus contribute to quantitatively assess the freedom of the press in a given country. Furthermore, they can be applied in wider contexts such as the evaluation of bias toward specific ethnic groups or specific accidents.

Keywords: News; Death; Suicide; Scaling laws; Censorship; Gender bias; Newspaper

1 Introduction

Violent death is a dramatic event that can be objectively quantified in terms of the number of lives lost. Words such as 'killed', 'dead', 'casualty', 'suicide' etc... can thus be used to

assess the presence of bias or censorship in reporting incidents involving specific locations, groups of people, or time.

Coverage of death in the mass media has been studied for the insight it offers on how societies perceive this dramatic event[1]; at the same time, selective death reporting can be used to shape and distort public opinion. A 6-months period study[2] of 1975 issues of the *Register Guard* and the *Standard Times* showed that newspapers tend to over-report more news-worthy death causes and overlook others that are considered less interesting by the public. The work considered the number of occurrences, the number of deaths reported, and the text surface area of the articles dealing with death and compared them with the statistical occurrences, finding that all forms of disease were under-reported whereas violent or catastrophic events were overrepresented. An even more biased selection along these lines was found analyzing the news selection by television and radio[3]. Studies on the presence of geographical bias in reporting accidental deaths have been conducted manually on small data sets: a 51-issues study (two months) of two German (*Frankfurter Allgemeine Zeitung* and *Süddeutsche Zeitung*) and two Australian newspapers (*The Australian* and *The Sydney Morning Herald*) [4] have shown a weak relationship with social proximity in the number of articles reporting deaths. Similarly, a 6-months study of the German tabloid *Die Abendzeitung*[5] reported evidence of a correlation between the number of accidental deaths and the length of the article as well as a negative correlation of the distance between the town where the newspaper was printed and the event reported. In this case, the analysis involved the study of the amount of printed space occupied by an article reporting of a fatal event. The authors did not find any significant correlation between the area taken by the article and the number of deaths reported[6] but showed an inverse relationship (linear in logarithmic scale) of the area of printed text normalized to the number of victims and the distance (in km) of the event from Munich, the town where the *Die Abendzeitung* was printed. They also found a minimum threshold of victims - increasing with distance - needed to elicit a report in the newspaper. They thus likened the response of the paper to the logarithmic response between stimulus strength and response intensity in human perception[7].

On the opposite, [8] analyzed the occurrence of pictures in *Time* and *Newsweek* magazines. The study involved ten issues in three years for each magazine, for a total of 60 issues. The work involved measuring the area occupied by the picture, determine the country of origin of the picture, and classifying its violent or non-violent nature. The study found that - although domestic pictures were more common than foreign ones (*Time=66.4*$\%$, *Newsweek=71.2%*) - images depicting death were printed more often when the event took place abroad (*Time*$_{foreign}$: 31%; *Newsweek*$_{foreign}$:25%) compared to when it occurred in domestic USA (*Time*$_{domestic}$: 12%; *Newsweek*$_{domestic}$: 13%).

A more recent study[9] showed that direct pictographic representations of death are very rare in western newspapers, amounting only to 4.5% of the articles (997 stories in a two-month time frame, 357 with pictures).

Most of these studies have been performed by manual scanning of the newspapers or magazines, therefore restricting the scope of the possible analysis. With the digitization of printed literature, it is possible to perform systematic analysis on the entire archives of newspapers, magazines, and books. A large-scale analysis of the number of occurrences of notable personalities in books has shown evidence of censorship of various individuals during the Nazi regime[10]. The books have been digitally converted via OCR (Optical

Character Recognition)[11] as part of the Google effort to digitize books and the occurrence of individual names has been compared with the lexicon of the various languages[12].

In this work, we have developed several algorithms that allow a systematic study of the occurrence of the articles involving fatal accidents and their gender/location. This allows to assess the distribution laws and determine the presence of bias and censorship in reporting. We have used daily newspapers since they have the advantage of addressing wider audiences with higher frequency and a longer temporal consistency, often covering events usually ignored by books.

We have applied these algorithms on the full historical archives of the three major Italian newspapers: *Il Corriere della Sera* (Milan, CDS, 1876-2017), *La Repubblica* (Rome, REP, 1985-2017), and *La Stampa* (Turin, STA, 1867-2005) and the major US newspaper: *The New York Times* (NYT, 1955-2017), searching for the articles containing selected keywords: *'killed'* (*'morti'* in the Italian language, the mixed-gender plural term, as well as the singular female and male gender, *'morta'* and *'morto'* respectively), *'suicide'* (*'suicidio'*). These words have been chosen for their high rate of occurrence (3.1 million articles out of 42 million comprising the newspaper archives) in respect to similar lemmas (dead, casualties...) and correspondence of usage since both are used to indicate violent deaths. Most importantly, in the case of the plural form they contain the number of people involved. Therefore, in addition to a simple word count, it is also possible to study the distribution of the number of people killed, the location, and the category of victims.

The rest of the paper is organized as follows: in the *Methods* section is present the description of the archives, the retrieval and data processing steps, and the various algorithms. In the *Results/Discussion* section are shown the results on the time behaviour, gender, censorship, the scaling law, and the last digit analysis. Perspectives and future work are discussed in the *Conclusion* section.

2 Methods

Figure 1 shows the steps taken to access, parse, and identify the articles, from the access of the online archives to the creation of the various data sets, one for each newspaper and lemma considered. All these tasks are accomplished using Python scripts. The access to the online archive of each newspaper takes a few days to complete over the full time span, since this is mediated by a web-based interface that usually returns ten articles at a time. The storage and handling of the data sets are performed using the Root[13] framework and require a few minutes to complete. This C++ based environment was developed at CERN to deal efficiently with high volumes ($\simeq$ Pbyte) of data produced from accelerator-space- and ground/underground-based detectors. As such, it is especially suited to create the ntuples (Ttrees in Root nomenclature) containing the data extracted from the newspaper articles. All the selections, histogramming, and fitting have also been performed in this environment.

A scheme of the various types of algorithms used to analyze the data sets and the main information they provide is shown in Figure 2.

2.1 Newspaper Archives

Only printed editions have been considered and online articles have been excluded for consistency with pre-internet years. The newspapers have freely accessible online archives that cover a major part of their printing time. Out of the $T \simeq 42$ million articles comprising

Newspaper	Date	T all entries	Suicide Suicidio	Morta Morto	K killed morti	N number present	G location present	N and G number and location	Archive
NYT	1852 2016	1.51e7	1.43e5		4.74e5	3.10e4	2.58e5	1.82e4	https://www.nytimes.com/search/
CDS	1876 2017	8.04e6	7.72e4	3.33e5 6.5e5	2.77e5	2.53e4	7.44e4	1.10e4	http://archivio.corriere.it/
REP	1985 2017	6.36e6	3.68e4	6.52e4 1.5e5	1.68e5	3.70e4	8.36e4	2.20e4	https://ricerca.repubblica.it/
STA	1867 2005	1.28e7	7.33e4	1.31e5 3.68e5	1.79e5	1.86e4	1.39e5	1.63e4	http://www.archiviolastampa.it/
Total		4.23e7	3.3e5	1.7e6	1.10e6	1.12e5	5.55e5	6.75e4	

Table 1 Size of the online newspaper archives consulted and of the resulting datasets according to the various selection criteria. The totality of the archives has been considered in this work.

the four newspaper online archives, all the articles which matched the considered lemmas (*'killed'*, *'morti'*, *'morta' 'morto' 'suicide'* (*'suicidio'* $\simeq 3.2$ million) have been extracted. See Table 1 for details on the data size of each newspaper. Details of the newspapers and archives are as follow:

1 *The New York Times* (NYT) was founded in September 1851, with the online archive (https://www.nytimes.com/search/) starting in January 1952. Due to strikes, it was not printed between: December 9, 1962 - March 31, 1963 (a western edition is present in the archive); September 17, 1965, - October 10, 1965 (an international edition was printed and is present in the archives); August 10, 1978 - November 5, 1978 (no editions present in the archives). In 2017, print circulation of the newspaper was 571,500 copies[1]. The newspaper began its publications with about 20,000 articles/year growing gradually to reach 130,000 articles in 2016.

2 *Il Corriere della Sera* (CDS). The first issue dates back to March 5th 1876 and is available in the online archive (http://archivio.corriere.it/). Originally an evening newspaper, it became a morning paper in 1888 and was issued twice a day since 1892 and up two to three times a day in the first part of the XXth Century, but it has been printed as a daily newspaper for several decades. Its offices were bombed on 14/2/1943. Following the liberation of Italy from Nazi occupation, the Committee of National Liberation (*Comitato di Liberazione Nazionale*) suspended its publications between 27/4 - 21/5 1945. It resumed publications under the name 'Corriere d'Informazione' and, from 1946, as *Il Nuovo Corriere della Sera*, with a one-page edition[14, 15]. CDS passes from $\simeq$7,000 articles/year of the first years to $\simeq$125,000 articles/year in 2017. The FAQ of the archive reports that it contains about 2.5 million of pages scanned. In December 2017 it printed 310,275 copies[2].

3 *La Repubblica* (REP) begun publications in January 1976, but the online archive (https://ricerca.repubblica.it/) starts on January 4th, 1984, with the first entry of "morti" occurring on March 4th, 1984. In December 2017 it printed 274,745 copies. REP starts with 24,000/year in 1985 to reach 377,000 articles in 2017.

4 *La Stampa* (STA). The digitalization of this archive (http://www.archiviolastampa.it/) was performed by the Committee for the Journalistic Information Library (*Comitato per la Biblioteca dell'Informazione Giornalistica CBDIG)*. The archive is released under a Creative Commons license and covers the period from the first issue, February 9th, 1967, when it was called *Gazzetta Piemontese* (*Piedmont Gazette*) to De-

[1]New York Times Company form 10-K, 2017.
[2] Data on number of printed copies retrieved from http://www.adsnotizie.it

cember 31st, 2005. Except for a few entries in 1882, all articles up to and included 1909 are referenced in image form with the title *'Notizia'* (*'News'*) and no further information on the contents of the article. Several entries after 1909 are stored in this way, thus reducing the overall time range of the dataset. The Fascist government halted publications of *La Stampa* in the month of October 1925 as a warning to all publishers. A few days after resuming publications (November 3rd 1925), Alfredo Frassati, Director of the newspaper for 25 years, resigned to be replaced by directors gradually more aligned with the government. Following the liberation of Italy from Nazi occupation, the Committee of National Liberation (*Comitato di Liberazione Nazionale)* suspended its publications between 28 April - 17 July 1945 (no entries in the database)[16, 17]. In December 2017 it printed 208,615 copies. STA increases from $\simeq$10,000 articles/year to a maximum of 500,000 articles/year in 2001, decreasing at $\simeq$ 200,000 in 2005, the last year present in the archive.

2.2 Lemmas considered

This work concerns the study of the occurrence of the following terms:

- **'Killed'** (in English) and **'morti'** (plural for Italian). Note that the literal translation of *killed* in Italian language is *uccisi*, but this word is less frequent than *morti* and usually (but not always) employed in association with murder. *Morti* is more often associated with violent accidental death and thus has a wider correspondence in usage with *killed*. Articles with the keywords *casualties/vittime*, occur with a lower frequency than *killed/morti* (10% to 30% less). The English term is more strongly associated with dead or wounded during armed conflicts and has peaks during major US wars (Civil, WW1, WW2, Korea, Vietnam etc...). In case of the Italian newspapers, the correlation is lower, since *vittime* is more often used for people killed in accidents, natural disasters, and so on.
- **'Suicide'** (in English) and **'suicidio'** (in Italian).
- **'morta'** and **'morto'** (in Italian), respectively feminine and masculine singular form of 'morti'.

Some methods discussed below can be applied to any lemma, although the distribution law and last digit analysis is usable in other 'quantitative' keywords such as *'dead', 'victims',* and the already mentioned *casualties*, will be the subject of future work.

2.3 Construction of the Data sets

1. Query and pre-processing. Each data set has been constructed by querying the web servers of the four newspapers with Python scripts that emulated manual user input (Figure 1). The input had to be configured for each archive in order to enter the desired keyword and iterate the requests over the time interval of the newspaper archive. The html pages received in reply to the query were then saved locally. Their overall number depends from the number of articles present in each page: we range from the $\simeq 58,000$ of NYT to the $\simeq$ 300 of CDS. As mentioned, this data acquisition / preprocessing phase over the newspaper archives takes 5 to 7 days to complete, depending on the newspaper considered[3].

We could not access the full articles since they are often present only in image form (especially in the years 1850-1910) and the necessary work and resources would have been

[3]We used the selenium (https://www.selenium.dev/) package with python bindings to automatize the interaction with the archives.

almost equivalent to those required for the digitization of the newspaper itself. Furthermore, the work would only have resulted in a marginally higher efficiency in event detection and would not have changed the results.

We define an 'event' as a newspaper article retrieved upon the query of the keyword killed/morti. In our data sets, each event contains the newspaper name, the date, and the page (not present in NYT) of the article, the text of the title, and - if present - a part of the text of the article.

2. Parse. The html pages are then subsequently parsed with Python regular expressions that extract the relevant article information (title, snippet, page number, date, etc...). This phase takes a few minutes to complete.

3. Filter. Sometimes the query only returns a link with no usable text: for instance a title can be empty (especially for issues of the XIX century), incomplete, or does not contain the requested word. These events have been discarded (in the case of STA, this restricts the database to the period 1910-2005 for most purposes).

4. Text Analysis. The events are then analyzed to find the number k of people killed. This is searched in numeric, text, and hybrid form. To reduce classification errors, the value of k is searched close to the keyword in the forms: *'k killed'*, *'killed k'*, *'k attribute killed'*, *'killed attribute k'*. Title and text are also parsed to search for the type of event (e.g. car, airplane crash, war, illness), and the people involved (e.g. children, women, ethnicity), that will be the subject of a following paper.

5. Geographical analysis. Several databases of world places have been used to associate the location event to its country of origin. Cases of homonymy (Florence, Paris, Cairo...) have been resolved assigning the location of the more famous ones at that time. For instance, 'Cairo' entries in the years 1861-1865 have been assigned to the US since the location appeared often during the American Civil War.

6. Exclusions. Duplicate entries, that is articles identical in title, text, and date are then removed, but different reports on the same event are considered as separate since it means that they have been considered worthy of more than one article. See Table 1 for the size of the datasets according to the various selections.

All events reporting deaths of animals (fish, herd, cows, etc.), usually associated with high k have been discarded.

From foreign events we also excluded all articles where the words 'Italian' or 'US'/'American' appear associated with foreign countries, thus removing foreign events where citizens from the corresponding newspaper-printing country were involved. This amounted to less than 1% in peacetime.

Once the parsing of the archives is completed, the remaining processing steps in this phase take a few minutes to execute and produce the database / root files for the subsequent analysis. This represents an increase of several orders of magnitute in speed in respect to any traditional, manual-scanning method which had to be necessarily constrained to a limited amount of time and newspaper issues.

2.4 Errors

Statistical errors are due to fluctuations in the number of events present in a given bin of a given selection. Fitting algorithms take into account the errors associated with each bin to calculate the errors of the fitted parameters.

Sources of systematic errors can be due to OCR (Optical Character Recognition) misidentification. This is more frequent for old issues where the quality of the scanned pages is lower and can result in a lower efficiency for the first years of the newspapers.

This can be assumed to be independent of the number, type, or location of people killed k so that the temporal behaviour and distribution laws should not be affected. See the Supplementary Information for a discussion on systematic errors.

3 Results and Discussion

The analysis of the data sets created above can yield information on how the newspapers consider the various events depending on geographic location, historical period, or gender. In this section, we describe the main algorithms employed and the results they provide. The various algorithms, the processing steps, and the key results derived, are also schematized in Figure 2.

3.1 Historical events

Figure 3 shows the yearly total number of articles, $T(t)$. Major historical events can also affect this value: for instance, it increases in NYT during the US Civil War due to more articles being published and decreases in CDS and STA during the two World Wars due to due to shortage of materials resulting in fewer pages being published.

In the same Figure the number of articles with the keyword 'killed', $K(t)$, is also shown. From it, we can derive the normalized fraction of 'killed' events in respect to the total: $R = K/T$ (shown in the same Figure), a value more affected by historical events.

For the events with a determined location, it is possible to separate domestic (K_d) from foreign (K_f) occurrences and assess how their relative importance evolved with time (Figure 3). NYT reporting on foreign deaths grows over time to become more frequent than domestic at the onset of WW1 and permanently from WW2 on. The Italian newspapers divide the reporting between domestic and foreign cases roughly equally, with CDS covering more foreign events in the more recent years and STA and REP the domestic cases. The other relevant features are the foreign peaks during the World Wars and the drop in the domestic deaths between 1923 and 1945 for STA and CDS, due to censorship from the Fascist government (see below).

In Figure 4 are shown the relative contributions of the various continents and their evolution over time, showing a gradual reduction of the coverage of European events and a growing importance of Asia after WW2. Some major occurrences are:

3.1.1 American Civil war (12/4/1861-13/5/1865)

During the American Civil war, the total number of NYT articles increases by 18.5% in respect to the interpolated values of T between 1861 and 1865. K increases to a maximum of $K_{1863}/K_{1860} = 2.80$ the pre-conflict value (the keyword *'casualties'* has an increase of $C_{1863}/C_{1860} = 4.23$). Conversely, the number of suicides drops to a minimum of $S_{1864}/K_{1860} = 0.36$.

3.1.2 World War 1 (28/7/1914-11/11/1918, Italy from 23/5/1915, US from 6/4/1917)

Censorship was very strong in all countries involved in both World Wars, with the removal of all information who could be beneficial to the enemy: letters, reporting of battles and defeats, casualties etc.[18].

In NYT the value of $R = K/T$ increases from $R_{1913} = 0.361 \pm 0.001$ to $R_{1914} = 0.457 \pm 0.001$. The domestic event ratio $R' = K_{dom}/K_{total}$ drops from $R'^{NYT}_{1913} = 0.56 \pm 0.02$ to $R'^{NYT}_{1914} = 0.39 \pm 0.01$, reaching a minimum of $R'^{NYT}_{1917} = 0.32 \pm 0.01$, when the US declared war to Germany.

In Italy, the shortage of resources resulted in a reduction of the number of pages of CDS and STA from 8 (two double sheets) to 4 (one double sheet of paper). Consequently, T decreased to $T^{CDS}_{1918}/T^{CDS}_{1915} = 0.48$ and $T^{STA}_{1918}/T^{STA}_{1915} = 0.67$. The effect of censorship is evident in STA: its value of R drops from $R_{1915} = 0.32 \pm 0.02$ to a minimum of $R_{1917} = 0.13 \pm 0.01$ (for CDS is more constant). In both newspapers, there is an even higher drop in domestic events (not necessarily only due to censorship but also to a lack of interest): from $R'^{STA}_{1914} = K_{dom}/K_{total} = 0.43 \pm 0.07$ to $R'^{STA}_{1917} = 0.26 \pm 0.09$. For CDS, R drops from $R'^{CDS}_{1914} = 0.51 \pm 0.06$ to $R'^{CDS}_{1917} = 0.29 \pm 0.04$.

During wars the rate of suicides is known to decrease: this phenomenon is usually explained by the higher sense of purpose during the bellic effort[19, 20, 21] and is found to occur both when one's country or other countries are at war. However, the decrease of suicides reporting by newspapers is more prompt and intense (dropping to 1/3 of the pre-war value) than that recorded by statistics. Since in both countries this occurred in 1914, before either country was at war we can conclude that this was not directly related to censorship. In Italy, the number of articles on suicides passes from $S^{STA}_{1913}/T^{STA}_{1913} = 0.043 \pm 0.002$ to $S^{STA}_{1914}/T^{STA}_{1914} = 0.032 \pm 0.002$ and reaches a minimum of $S^{STA}_{1917}/T^{STA}_{1917} = 0.012 \pm 0.001$ before returning to $S^{STA}_{1919}/T^{STA}_{1919} = 0.025 \pm 0.002$. A similar behaviour is found for CDS (and NYT): from $S^{CDS}_{1913} = 1.34 \pm 0.06$ ($S^{NYT}_{1913} = 1.18 \pm 0.04$) to $S^{CDS}_{1914} = 0.73 \pm 0.04$ ($S^{NYT}_{1914} = 0.92 \pm 0.04$) at the beginning of WW1 to a minimum $S^{CDS}_{1916} = 0.45 \pm 0.03$ ($S^{NYT}_{1916} = 0.39 \pm 0.02$). The drop before the US or Italy entered the war suggests to attribute the lack of suicide reporting to a reduced interest by the editorial rooms rather than to censorship.

3.1.3 Fascist government in Italy (31/10/1922 - 25/7/1943)

Different is the case during the Fascist government in Italy. The Italian government of the time exercised a strong censorship on printed press and radio. On 14/7/1924 a *circolare* (note) from the then Minister of Interior Federzoni allows the sequestering of copies of newspapers to 'prevent stirring up public opinion'. On 31/12/1924 all newspapers are sequestered and the directors replaced with ones affiliated with the regime. In October 1926, several daily newspapers were closed until the end of WW2. Among them *L'Unità*, *L'Avanti!* and *L'Ora*[15, 14].

Government censorship aimed to present an efficient state and thus had to remove all negative news. Censorship involved all media of the time: radio, theater, movies, books, and newspapers. Authors, especially those of Hebraic origin but also those who were against the regime for political reasons, fell in disfavour. This 'targeted' censorship was similar to what occurred in Germany and reported in[10], where prominent individuals were mentioned to a greater or lesser extent according to their race or standing in respect to the Nazi government.

On a wider scale, government guidelines[16, 17] to newspapers required that crime reporting had to be compressed in a few lines, and suicides had to be ignored, with the result that articles involving domestic deaths and accidents almost disappeared from newspapers. With the datasets of CDS and STA it is possible to quantify the overall effect of Fascist censorship in the reporting of violent deaths [22, 23].

As a result, even though the value of $R = K/T$ remains more or less constant, in 1923 R' drops from $R'^{CDS}_{1922} = 0.56 \pm 0.03$ ($R'^{STA}_{1922} = 0.64 \pm 0.05$) to $R'^{CDS}_{1923} = 0.44 \pm 0.02$ ($R'^{STA}_{1922} = 0.46 \pm 0.05$) reaching a minimum of $R'^{CDS}_{1936} = 0.119 \pm 0.005$ ($R'^{STA}_{1937} = 0.31 \pm 0.01$) (Figure 3). These values come back to pre-dictatorship values of $R'^{CDS}_{1946} = 0.60 \pm 0.03$ ($R'^{STA}_{1946} = 0.68 \pm 0.06$) immediately after the war when both newspapers had their publications halted and their directors were replaced between April and May (July for STA) 1945.

We estimated the amount of censorship for articles with *morti* interpolating the value of R' between 1922 and 1946: between 1923 and 1943 there were 2,800±200 domestic articles with at least two casualties missing for CDS and 2,900±300 for STA. In both cases, they amount to 75% of the articles featuring domestic deaths printed in the same period.

A similar analysis on the $k = 1$ events (*morto, morta*, see Figure 5) yields about $22,000 \pm 1300$ articles censored by STA (60% of those printed in the period considered) and $29,000 \pm 500$ by CDS (20% of those printed in the period considered).

In the case of suicides (another forbidden topic during the regime[15]), Figure 6 gives $S_{missing} = 17,000 \pm 600$ ($S_{missing}/S_{present} = 2.8$) for CDS and 4,100±300 ($S_{missing}/S_{present} = 1.1$) for STA, in contrast with the growing rate of suicides at the time[20, 24].

Overall we estimate that CDS censored $41,800 \pm 1000$ articles and STA $36,000 \pm 1,900$ during this period, with an average of $1990 \pm 50/year$ for CDS and $1700 \pm 90/year$ for STA.

3.1.4 World War 2 (1/9/1939 - 2/9/1945, Italy from 10/6/1940 as part of the Axis and from 8/9/1943 as part of the Allies, USA from 7/12/1941)

Censorship in US during WW2 relied mostly upon the self-censorship of news outlets: forbidden topics included weather and crop reports, correspondence, travel schedules and naturally troop movements[25]. The Office of Censorship released on January 15, 1942, the Code of Wartime Practices for American Broadcasters and the Code of Wartime Practices for the American Press. The publication of any pictures depicting US soldiers killed in combat was forbidden until September 1943, when the capitulation of Italy might have induced in the general public the idea that the war was to end soon[18].

The total number of articles of NYT remains essentially unchanged, with R increasing only in 1940, from $R_{1938} = 2.02 \pm 0.04\%$ to $R_{1940} = 2.50 \pm 0.04\%$. This is due to an increased amount of reporting of casualties prior to the entrance in the conflict. This is confirmed by the value of R', from $R'_{1938} = K_{foreign}/K_{domestic} = 0.63 \pm 0.01$ to a maximum of $R'_{1940} = K_{foreign}/K_{domestic} = 0.76 \pm 0.01$, again before Pearl Harbor. Afterward, since US soil was not attacked, the amount of reporting of domestic events was not affected so we can conclude it was not affected by censorship.

The effect of WW2 in Italy is mostly evident in the sharp drop in T and K in the later years of the war. This was both due to paper shortage and the bombings on Milano and Torino, where the newspapers were printed. After the armistice with the Allied forces (8/9/1943), Italy was split between the South, controlled by the Allies and *La Repubblica di Salò* in the North. After a short period free from Fascist government, CDS and STA are then aligned to the Nazi-controlled government of the North so the definition of 'domestic' and 'foreign' becomes fuzzier and the ratio R' increases.

3.1.5 Recent decades

If we consider the reporting of deadly accidents in the last decades we see that it is constant or decreasing: in the NYT, after doubling between 1960 and 1969, R^{NYT} remains - with fluctuations - constant up to 2017.

The Ratio R for STA decreases gradually after 1954, with $dR/dt_{STA\,1954-2005} = (-0.1 \pm 0.02)\%/year$.

Also CDS exhibits a similar but stronger decrease: $dR/dt_{CDS\,1954-2005} = (-0.3 \pm 0.02)\%/year$.

For REP, the amount of articles reporting violent death K increases with time but its percentage after 1995 decreases by $dR/dt_{1995-2005} = (-0.2 \pm 0.04)\%/year$. We note that this trend of decreasing coverage given to violent events in all three Italian newspapers is opposite to the growth of the perceived threat of violence in Italy[26], so this phenomenon cannot be ascribed directly to press coverage.

3.2 Gender Bias

In Italian newspapers, where the language allows to distinguish gender, we have also queried the databases for the term *'morto'/'morta'* (M), respectively the singular ($k = 1$) masculine, and feminine form in the Italian language. *'Morte'*, the feminine plural could not be used since it also means 'death' in Italian and it would be difficult to semantically separate the two meanings. Besides, if both males and females are involved the term *'morti'* is used in Italian language, making the lemma *'morte'* of little use.

From Figure 5 we see that the amount of reporting of female deaths (*morta*) is only 30% of all $k = 1$ deaths. This has to be compared with the fatal accident standardized death rate (in 2005) in Italy of 36.1 (male) and 19.2 (female) per 100,000 deaths [27], and the probability for a 15-year old individual to die within 45 years, before reaching the age of 60 (45q15)[28] in 2010 of 7.9% for men and 4.1% for women.

Therefore, even accounting for the fact that male violent deaths are more frequent than female ones[4] and that these events would be more likely to be reported in the news, this still hints to some amount of gender bias in reporting. The female/male ratio of $\simeq 30\%$ is present in all newspapers and roughly constant over the years, with only REP showing an increase of reporting of female deaths of about $3\%/year$, from 22.6% in 1985 to 37.65% of 2007.

3.3 Scaling laws

The analysis of keywords associated with the number of persons involved allows to build the distribution function of the number of articles, N_k, reporting k people killed. As shown in Figure 7, the distributions for all four newspapers considered can be described by a single power-law:

$$N_k = N(k) = A \cdot k^{-\gamma} \tag{1}$$

in the range $2 \leq k \leq 10^6$. The sharp peaks in N_k for values of k that are multiples of 10, 100, 1000... are due to the rounding in excess to the nearest multiple of a power of 10 of the number of people reported (see below).

[4]For instance, in US the number of fatal occupational injuries for US in 2016 involved women in 7% and men in 93% of cases, for an almost equal number of worked hours (43% for women and 57% for men)[29].

The Minuit[30] package (now in its second release, Minuit2) has been used to perform the fits.

Fitting of the power laws has been analogous to the methods that we used in the fitting of cosmic ray spectra of the PAMELA space-borne detector[31], [32], [33] (see also ext. data therein). Tests with varying bin sizes have shown no significant change in the fitting results and values of γ.

See Section 7 for the mathematical details; a discussion of the fitting methods and error systematic can be found in the Supplementary Information.

Power-law statistics has been found to describe the distribution of various natural phenomena, e.g. earthquakes[34], forest fires[35], [36], the cluster size of tropical trees[37], and is usually thought to arise through positive growth feedback [38], [39].

Galactic cosmic ray spectra also follow a power law, as a result of the statistical process of acceleration. Changes in the spectral indexes show the presence of additional sources or production phenomena[32], [33].

Power law distributions are also encountered in many human-related activities[40], from language distribution (Zipf law[41], [42]), the number of casualties in wars[43], [44], and ethnic violence [45]. Also in these cases, they have been shown to arise through a 'winner takes all' type of a competitive network where a few elements grow to acquire a very large size [46], [47].

In newspapers, the distribution N_k can be explained as the result of two main phenomena:

- The convolution of various violent events and accidents. Road, train, air accidents, natural disasters and catastrophes have each their frequency and probability distribution, usually unknown but decreasing as k increases.
  Articles with $k \simeq> 1000$ often do not describe a specific accident, but rather summarize global phenomena such as war, illness deaths (cancer, heart attack...) or automobile accidents per year, etc... .
- The selection by the newsroom. The publishing criteria can change with time, location or censoring: events with higher k will have a higher probability of being selected for their importance. Conversely, foreign events, occurring in countries physically or socially far from the country where the newspaper is printed, will be more likely to be ignored, especially for low k.

Both phenomena can be approximated by a power law, with probability $P \propto k^{-\gamma_1}$ for an event to involve k casualties, and a probability (or efficiency) $\epsilon \propto k^{+\gamma_2}$ for the event to be picked by the newsroom. The overall probability is then $P_{tot} = \epsilon \cdot P \propto k^{-\gamma_1+\gamma_2}$ with $\gamma = \gamma_1 + \gamma_2$.

The values of γ of the four newspapers ($\gamma_{NYT} = 1.44 \pm 0.06$, $\gamma_{CDS} = 1.61 \pm 0.01$, $\gamma_{REP} = 1.42 \pm 0.07$, $\gamma_{STA} = 1.52 \pm 0.09$, Table 2) show that the editorial processes are similar in Italy/US and converge to a narrow range of values, although the differences among them reflect the different emphasis to high/low k events in the newspapers. A 'steep' spectrum, with a high γ, results in a higher number of articles with low k and vice-versa. Defining $W = N_H/N_L = N_{k>10}/N_{2\leq k\leq 10}$ we see (Table 2) that it ranges between $W_{CDS} = 0.57$ (more articles with low k) and $W_{REP} = 1.00$ (fewer articles with low k) with $W_{STA} = 0.72$ and $W_{NYT} = 0.93$ having intermediate values (see Section 7 and Figure 12 for a calculation of the value of W as a function of γ).

The trend of the spectral index can be used to estimate the state of belligerence reported by the newspapers: the running average (current and four preceding years) of $\gamma(t)$ (Figure

Newspaper	Loc.	γ	χ^2/NDF	N_H/N_L
NYT	All	1.44± 0.01	9	0.93 ±0.01
CDS	All	1.61± 0.01	7	0.57 ±0.01
REP	All	1.42e± 0.01	16	1.00 ±0.01
STA	All	1.52± 0.01	9	0.72 ±0.01

Table 2 Spectral index γ for a single power-law fit ($2 \leq k \leq 10^6$) on the full datasets of the four newspapers.

Newspaper	Loc.	γ_L	γ_H	$\Delta\gamma$ $\gamma_L - \gamma_H$	χ^2/NDF	M (Excess/Defect)
NYT	Dom.	1.37e+00± 3e-03	1.78 ± 1e-02	-0.417± 1e-02	3e+01	4%
NYT	For.	9.90e-01± 3e-03	1.66 ± 5e-03	-0.672± 5e-03	6e+01	-98%
CDS	Dom.	2.08e+00± 3e-03	1.81 ± 2e-02	0.277± 2e-02	2e+01	71%
CDS	For.	5.02e-01± 3e-03	1.59 ± 5e-03	-1.09± 6e-03	1e+02	-122%
REP	Dom.	1.69e+00± 3e-03	1.54 ± 6e-03	0.143± 7e-03	2e+02	16%
REP	For.	9.66e-01± 3e-03	1.48 ± 4e-03	-0.511± 5e-03	1e+02	-91%
STA	Dom.	1.95e+00± 2e-03	1.72 ± 1e-02	0.2311± 1e-02	4e+01	47%
STA	For.	9.23e-01± 3e-03	1.54 ± 5e-03	-0.615± 6e-03	9e+01	-94%

Table 3 Spectral index for the four newspapers separating domestic and foreign datasets. $\gamma_L = \gamma(2 \leq k \leq 10)$, $\gamma_H = \gamma(10 \leq k \leq 10^6)$. A negative $\Delta\gamma = \gamma_L - \gamma_H$ implies a lack of events for $2 \leq k \leq 10$ events, and vice-versa. The higher the absolute value the higher the excess or defect of events. M is the excess (+)/defect (-) of articles with respect to what expected from a pure power-law (see text).

8) decreases in wartime due to the higher abundance of high k events results in a flatter distribution. Vice-versa in peacetime, when the distributions are dominated by low k events, γ increases due to a steeper distribution. Thus, local minima in $\gamma(t)$ are present in NYT during the US Civil War ($\gamma = 1.21$), the two World Wars ($\gamma_{WW1} = 1.26$, $\gamma_{WW2} = 1.30$) and the Vietnam War ($\gamma = 1.31$). Also, CDS and STA show similar local minima during WW1 ($\gamma^{CDS} \simeq 1.2$ and $\gamma^{STA} = 0.8$) and WW2 ($\gamma^{STA} = 1.76$, $\gamma^{CDS} = 1.91$), followed by a sharp increase in STA (and more gradual in CDS) after 1945. It is also interesting to note that - notwithstanding the differences in γ - the trends of the spectral indexes of the newspapers are in good agreement among each other. This suggests that they all tend to react similarly to the conflicts occurring in the world (and vice-versa a discrepancy in the trend would indicate the presence of censorship).

3.4 Geographical bias

In our case, using the articles where a location could be determined, we built distribution laws for domestic and foreign events. In Figure 7 it is possible to see that the latter have a strong spectral break for $2 \leq k \leq 10$ (absent or less prominent in the former). Fitting the distribution function with two power laws $\gamma_H = \gamma(k > 10)$ and $\gamma_L = \gamma(2 \leq k \leq 10)$ we see that $\Delta\gamma = \gamma_L - \gamma_H$ is always negative for foreign events, ranging from $\Delta\gamma_{REP} = -0.511 \pm 0.005$ to $\Delta\gamma_{CDS} = -1.09 \pm 0.01$. For Italian newspapers featuring domestic events, the distributions are closer to a pure power law (highest $\Delta\gamma = +0.28 \pm 0.02$ for CDS). In all cases $\Delta\gamma$ is positive, implying that a higher emphasis is given to low k events. NYT has a smaller discrepancy between foreign ($\Delta\gamma = -0.672 \pm 0.003$) and domestic events ($\Delta\gamma = -0.42 \pm 0.01$). A plot of the value of γ vs $\Delta\gamma$ (Figure 9) shows that the foreign and domestic categories are clearly separated in all four newspapers (Table 3). This suggests a difference in the editorial behaviour due to a lack of press coverage of accidents involving a small number of persons in foreign countries, considered to be not interesting enough to be reported in the press.

Newspaper	All World δW $(\%/1000km)$	Europe / America δW $(\%/1000km)$
NYT	2.70±0.3	12±2
CDS	14.4±0.9	30±3
REP	11.5±1	40±9
STA	17±1	35±4

Table 4 Slope of the linear fit of the ratio $N_{2\leq k\leq 10}/N_{k>10}$ as a function of the distance between Rome/Washington and the various world countries.

An estimation of the under- or over-reporting of low k events can be provided by the extrapolation of γ_H to $2 \leq k \leq 10$:

$$M = \frac{N_L - \int_2^{10} \alpha_H k^{-\gamma_H} dk}{N_L} \tag{2}$$

with $N_L = \Sigma_2^{10} N_k$ and α_H coming from the power law fit of $k \geq 10$. Thus, M is the fraction of events with $2 \leq k \leq 10$ missing ($M < 0$) or exceeding ($M > 0$) the value expected from a pure power law distribution ($M = 0$). All four newspapers have $M_{for} \simeq -100\%$ in the foreign case, meaning that the editorial room decides to print only one event out of two if it involves ten or fewer casualties in a foreign country. We also note that Italian newspapers tend to print more news of domestic events (from $M_{dom} = 16\%$ of REP to $M_{dom} = 71\%$ of CDS) than what expected from a pure power-law distribution. This can be attributed in part to a large domestic and local news section. Overall CDS has the largest difference in dealing with foreign and domestic events ($M_{for}^{CDS} = -122\%$ - $M_{dom}^{CDS} = +71\%$) and the NYT the smallest ($M_{for}^{NYT} = -98\%$ - $M_{dom}^{NYT} = +4\%$). See Section 7 and Figure 13 for a calculation of the value of M as a function of $\delta\gamma$ and γ_H.

In many nations, there are too few events reported to perform an acceptable fit with a power law, therefore - for countries having at least 30 entries in a given newspaper - we used the ratio the $W = N_{2\leq k\leq 10}/N_{k>10}$ as an indicator of the intrinsic importance assigned to a given country.

In Figure 10, W_i is plotted as a function of the distance D_i between the capital of the country i and Rome/Washington, according to the newspaper. It is possible to see how the value of W tends to increase with the distance (fewer events with low k and more with high k). The geographical bias appears to be stronger in Italian newspapers: a linear fit (Table 4) shows that the slopes are similar in Italian newspapers and higher by a factor $\simeq 5$ compared to NYT, a sign of an higher internationalization of the US paper: $dW/dx_{CDS} = (14 \pm 1)\%/1000\ km$; $dW/dx_{REP} = (11 \pm 1)\%/1000\ km$ and $dW/dx_{STA} = (17 \pm 1)\%/1000\ km$, $dW/dx_{NYT} = (2.7 \pm 0.3)\%/1000\ km$.

Social proximity effects also play a role in defining the values of the various countries: as also visible in Figure 10, American countries have lower W than equally distant Asian and African ones.

If we limit the fit to countries in Europe (for Italian newspapers) and in America (for NYT) we find that the slopes are higher by a factor 2 to 4 than for those considering all world nations (Table 4) a sign that geographical distance plays a stronger role for countries that are socially closed to either Italy or US (although the different geography of the American continent plays a role in the different behaviour).

3.5 Editorial rounding by excess of casualties as an additional tool to detect censorship

Newspapers often round up the number of casualties reported: this can be due to lack of knowledge, to simplify the headline, or to purposely increase emphasis to attract the attention of the reader. In the absence of tampering, we would expect the least significant digit of k, l_k to follow a uniform distribution ($P(l_k) = 1/10$). However, in Figure 11, which shows the distribution l_k for $10 < k \leq 100$, it is possible to see how the values '0' and '5' are overabundant ('5' only in the Italian papers) and the others under-abundant in respect to the flat distribution expectation. The number of defects in the digits '6' to '9' are close to the excess of '0' (and similarly for '1' to '4' with '5'), confirming the artificial nature of the reported number of casualties. Overall, Italian newspapers have a value of $l_0 + l_5 = 40 - 46\%$ and NYT has a value of 32% (with respect to the 20% expected for the sum of the $k_0 + k_5$ bins). All distributions (see Table T1 in the suppl. mat.) are incompatible with the null hypothesis of the flat distribution with $p > 0.01$.

This distribution is found in all newspapers and historical periods with one notable exception: during the Fascist government, the domestic distributions of STA and CDS do not exhibit the peaks for k_0 and k_5, still present in the corresponding foreign distributions of the same period. A χ^2 test allows us to reject the hypothesis that foreign and domestic histograms of Italian newspapers follow the same distributions with $p > 0.01$. Furthermore, the domestic distributions of STA and CDS are the only ones that are not incompatible with the equiprobable one (Figure S3 and Table T1 in the Supplementary Information). This implies that during the Fascist regime the editorial practice was to report more faithfully the number of domestic casualties as an additional way of suppressing these events and at the same time exaggerating the number of foreign casualties.

This phenomenon is similar but specular to Benford's law[48, 49], which describes the statistical distribution of the most significant digit in several natural and man-made datasets[50]. Benford's law has been used to determine the presence of accounting[51] or election fraud[52], since values that are artificially altered do not follow it.

4 Conclusions

In this work we have developed a series of techniques to automatically scan the complete historical databases of daily newspapers for the occurrence of specific keywords related to accidents and death. We also devised various algorithms to analyze the dependence of these keywords from the geographical location and historical period. Over traditional analysis, consisting of manual scanning of the newspaper articles, these tools offer the advantage of being automatic and thus being applicable on larger datasets spanning longer time periods. Indeed, these tools are suited for historical analysis to evaluate quantitatively the presence of bias or censorship in a given publication and its variation over time and political environment. This paper considered printed daily newspapers, but these techniques can be used also on online publications, news outlets, etc... Although the usage is limited only to articles with quantitative keywords, where to the event (accident, casualty, death...) is associated the number of persons involved (The simpler word occurrence methods can be used on a wider word set), this type of analysis is complementary to the assessment of 'fake news' since has the advantage of being automatic and operating on large data sets. These tools can also contribute to the assessment of the freedom of the press in a given country.

Future work will extend the application of these tools to other lemmas such as *casualties, wounded, victims*. The analysis will be applied to differences in reporting between the type

of accidents, both of man-made origin (e.g. train/airplane/ship) and natural calamities. Also the structure of reporting as a function of the day of the week and the page location will be considered. On a larger scope. also other newspapers, magazines, online publications will be considered, extending the analysis to look for the presence and evolution of ethnic or national bias. A more long-term goal can also be the use of speech recognition methods to study the occurrence of these lemmas on radio and TV.

5 Abbreviations

- CDS - Il Corriere della Sera
- NYT - The New York Times
- OCR - Optical Character Recognition
- REP - La Repubblica
- STA - La Stampa
- WW1 - World War 1
- WW2 - World War 2

6 Declarations

6.1 Availability of data and material

The newspaper archives are accessible at these locations:

New York Times: https://www.nytimes.com/search/

Il Corriere della Sera http://archivio.corriere.it/

La Repubblica: https://ricerca.repubblica.it/

La Stampa: http://www.archiviolastampa.it/

ROOT Ntuples with processed data used in the analysis will be made available upon reasonable request.

6.2 Competing interests

None

6.3 Funding

None

6.4 Authors' contributions

Not Applicable

6.5 Authors' information

Marco Casolino's research topics are mostly focused on the study of cosmic rays and radiation environment with the construction and launch of space-borne detectors, both on satellites and International Space Stations. Among the satellite-borne detectors he was involved in the Nina and Mita instruments to study the MeV-GeV component of cosmic rays, and in the Pamela magnetic spectrometer to perform the first detailed measurement of the antiparticle component (positrons, antiprotons) in galactic cosmic rays, of relevance for astrophysics and indirect dark matter search. On the Space station Mir and International he was involved in the study of the light flash phenomenon observed by astronauts and in the determination of the radiation environment in low Earth orbit. He has been PI of several instruments onboard the International Space Station, the most recent of which is

the Mini-EUSO telescope, observing the night-time Ultraviolet emissions of terrestrial, atmospheric, and cosmic origin. He was also involved in a technological transfer project to apply detector technology from space to the realization of a detector for radiation in food after the Fukushima accident. This work utilized many of the data processing techniques developed for the analysis of the data coming from these instruments.

6.6 Acknowledgments

The author is grateful to Dr. A. Truzzi for the useful and stimulating discussions during the preparation of this work and Drs U. T. Casolino, W. Husein, O. Larsson, L. Marcelli, L. Sorge and M. Piersanti for reviewing the manuscript.

The author also wishes to thank the four newspapers considered (*The New York Times, Il Corriere della Sera, La Repubblica, La Stampa*) for having their historical archives freely available for consultation: without these resources the paper could not have been written.

7 Appendix. Estimation of the missing events

For a power-law distribution of the number of articles $N(k)$ reporting that k people have been killed:

$$N(k) = \alpha k^{-\gamma} \tag{3}$$

the total number of articles with two or more people killed is:

$$N_{tot} = \frac{\alpha}{\gamma - 1} 2^{1-\gamma} \quad \gamma > 1 \tag{4}$$

for a given N_{tot} we have therefore that:

$$N(k) = \frac{N_{tot}(\gamma - 1)}{2^{1-\gamma}} k^{-\gamma} \tag{5}$$

In a single power law, the ratio W of articles with less (N_L) and more (N_H) than 10 people dead, is:

$$W = \frac{10^{1-\gamma}}{2^{1-\gamma} - 10^{1-\gamma}} = \frac{1}{5^{\gamma-1} - 1} \tag{6}$$

A high value of γ implies an emphasis on articles with low k and vice-versa. For $\gamma = 1.43$ we have an equal number of articles with $2 \leq k \leq 10$ and $k > 10$ (Figure 12).

For a broken power law distribution:

$$N(k) = \begin{cases} \alpha k^{-\gamma_L} & 2 \leq k \leq 10 \\ \beta k^{-\gamma_H} & k > 10 \end{cases} \tag{7}$$

so

$$N_L = \frac{\alpha}{1 - \gamma_L} \left(10^{1-\gamma_L} - 2^{1-\gamma_L}\right) \quad 2 \leq k \leq 10 \tag{8}$$

$$N_H = -\frac{-\beta}{1 - \gamma_H} \, 10^{1-\gamma_H} \quad k > 10 \tag{9}$$

$$M = 1 - \frac{\int_2^{10} \beta k^{-\gamma_H} dk}{\int_2^{10} \alpha k^{-\gamma_L} dk} = 1 - \frac{1-\gamma_L}{1-\gamma_H} 10^{\gamma_H - \gamma_L} \frac{10^{1-\gamma_H} - 2^{1-\gamma_H}}{10^{1-\gamma_L} - 2^{1-\gamma_L}} \tag{10}$$

The excess ($M > 0$) or defect ($M > 0$) in respect to a pure power law ($M = 0$) thus depends on the two values γ_H and γ_L. They are shown in Figure 13.

References

1. Ariès P, Ranum P. Western Attitudes toward Death: From the Middle Ages to the Present. Western Attitudes toward Death. Johns Hopkins University Press; 1975. Available from: https://books.google.it/books?id=3sZJN3wojesC.
2. Combs B, P S. Newspaper coverage of causes of death. Journalism Quarterly's. 1979;61:837–843,849.
3. Altheide DL. Creating Reality: How TV News Distorts Events. A SageMark edition. Sage Publications; 1977. Available from: https://books.google.it/books?id=ml21yAEACAAJ.
4. Hanusch F. VALUING THOSE CLOSE TO US. Journalism Studies. 2008;9(3):341–356. Available from: https://doi.org/10.1080/14616700801997281.
5. Burdach KJ. Reporting on Deaths: The Perspective Coverage of Accident News in a German Tabloid. European Journal of Communication. 1988;3(1):81–89. Available from: https://doi.org/10.1177/0267323188003001005.
6. Stigler SM. Francis Galton's Account of the Invention of Correlation. Statist Sci. 1989 05;4(2):73–79. Available from: https://doi.org/10.1214/ss/1177012580.
7. Dember WN, Warm JS. Psycology of Perception. New York: Holt, Rinehart and Winston; 1979.
8. jen Tsang K. News Photos in Time and Newsweek. Journalism Quarterly's. 1984;56(723):578–584.
9. Hanusch F. Graphic death in the news media: present or absent? Mortality. 2008;13(4):301–317. Available from: https://doi.org/10.1080/13576270802383840.
10. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, , et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2011;331(6014):176–182. Available from: http://science.sciencemag.org/content/331/6014/176.
11. Smith R, Antonova D, Lee DS. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR. In: Proceedings of the International Workshop on Multilingual OCR. MOCR '09. New York, NY, USA: Association for Computing Machinery; 2009. Available from: https://doi.org/10.1145/1577802.1577804.
12. Niyogi P. The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar. Springer US; 1998.
13. Brun R, Rademakers F. ROOT — An object oriented data analysis framework. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 1997;389(1):81 – 86. New Computing Techniques in Physics Research V. Available from: http://www.sciencedirect.com/science/article/pii/S016890029700048X.
14. Smith DM. Storia di cento anni di vita italiana visti attraverso il Corriere della sera. Rizzoli; 1978.
15. Melograni P. Il Corriere della sera (1919-1943). Cappelli Editore; 1965.
16. Nicola Tranfaglia ML Paolo Murialdi. La stampa italiana nell'età fascista. Editori Laterza; 1980.
17. Murialdi P. Storia del giornalismo italiano. Il Mulino; 2014.
18. James D Ciment TR. The Home Front Encyclopedia: United States, Britain, and Canada in World Wars I and II. vol. vol 1-3. 1st ed. ABC-CLIO; 2006.
19. Rojcewicz SJ. War and Suicide. Suicide and Life-Threatening Behavior;1(1):46–54. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1943-278X.1971.tb00598.x.
20. Somogyi S. Il suicidio in Italia 1864-1962. Analisi statistica. Milano: Giuffrè edit; 1967.
21. O'Malley P. SUICIDE AND WAR: A Case Study and Theoretical Appraisal. The British Journal of Criminology. 1975;15(4):348–359. Available from: http://www.jstor.org/stable/23636204.
22. Felice RD. Mussolini e il fascismo. Mussolini il duce. Lo stato totalitario 1936-1940. vol. Vol. 5. Einaudi; 1996.
23. Aquarone A. L'organizzazione dello Stato totalitario. Einaudi; 1995.
24. Alessandri G. Il suicidio in Italia. Aggiornamenti sociali. 1971;1(22):45–56.
25. Sweeney MS. Secrets of Victory: The Office of Censorship and the American Press and Radio in World War II. 1st ed. The University of North Carolina Press; 2001.
26. ISTAT. La percezione della sicurezza. Statistical Report. 2018;Available from: https://www.istat.it/it/archivio/217502.
27. Eurostat;. [Online; accessed September 2018]. https://ec.europa.eu/eurostat/web/health/causes-death.
28. Rajaratnam JK, Marcus JR, Levin-Rector A, Chalupka AN, Wang H, Dwyer L, et al. Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: a systematic analysis. The Lancet. 2010 May;375(9727):1704–1720. Available from: https://doi.org/10.1016/S0140-6736(10)60517-X.
29. of labor statistics USB;. [Online; accessed September 2018]. https://www.bls.gov/iif/oshwc/cfoi/cfch0015.pdf.
30. James F. MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1. CERN-D-506. 1994;.
31. Adriani O, Barbarino GC, Bazilevskaya GA, Bellotti R, Boezio M, Bogomolov EA, et al. The PAMELA Mission: Heralding a new era in precision cosmic ray physics. Physics Reports. 2014;544(4):323 – 370. The PAMELA Mission: Heralding a new era in precision cosmic ray physics.
32. Adriani O, Barbarino GC, Bazilevskaya GA, Bellotti R, Boezio M, Bogomolov EA, et al. PAMELA Measurements of Cosmic-Ray Proton and Helium Spectra. Science. 2011 Apr;332:69–.
33. Adriani O, Barbarino GC, Bazilevskaya GA, Bellotti R, Boezio M, Bogomolov EA, et al. An anomalous positron abundance in cosmic rays with energies 1.5-100GeV. Nature. 2009 Apr;458:607–609.

34. Bak P, Tang C. Earthquakes as a self-organized critical phenomenon. Journal of Geophysical Research: Solid Earth;94(B11):15635–15637. Available from: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JB094iB11p15635.
35. Malamud BD, Morein G, Turcotte DL. Forest Fires: An Example of Self-Organized Critical Behavior. Science. 1998;281(5384):1840–1842. Available from: http://science.sciencemag.org/content/281/5384/1840.
36. Telesca L, Amatucci G, Lasaponara R, Lovallo M, Rodrigues MJ. Space time fractal properties of the forest-fire series in central Italy. Communications in Nonlinear Science and Numerical Simulations. 2007 Oct;12:1326–1333.
37. Condit R, Ashton PS, Baker P, Bunyavejchewin S, Gunatilleke S, Gunatilleke N, et al. Spatial Patterns in the Distribution of Tropical Tree Species. Science. 2000;288(5470):1414–1418. Available from: http://science.sciencemag.org/content/288/5470/1414.
38. Scanlon TM, Caylor KK, Levin SA, Rodriguez-Iturbe I. Positive feedbacks promote power-law clustering of Kalahari vegetation. Nature. 2007 Sep;449:209 EP –. Available from: http://dx.doi.org/10.1038/nature06060.
39. Yakovlev G, Newman WI, Turcotte DL, Gabrielov A. An inverse cascade model for self-organized complexity and natural hazards. Geophysical Journal International. 2005 Nov;163:433–442.
40. Clauset A, Shalizi CR, Newman MEJ. Power-Law Distributions in Empirical Data. SIAM Review. 2009;51(4):661–703. Available from: https://doi.org/10.1137/070710111.
41. Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005;46(5):323–351. Available from: https://doi.org/10.1080/00107510500052444.
42. Moreno-Sanchez I, Font-Clos F, Corral A. Large-Scale Analysis of Zipf's Law in English Texts. PLOS ONE. 2016 01;11(1):1–19. Available from: https://doi.org/10.1371/journal.pone.0147073.
43. Richardson LF. Variation of the Frequency of Fatal Quarrels with Magnitude. Journal of the American Statistical Association. 1948;43(244):523–546. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1948.10483278.
44. Cederman LE. Modeling the Size of Wars: From Billiard Balls to Sandpiles. American Political Science Review. 2003;97(1):135–150.
45. Lim M, Metzler R, Bar-Yam Y. Global Pattern Formation and Ethnic/Cultural Violence. Science. 2007;317(5844):1540–1544. Available from: http://science.sciencemag.org/content/317/5844/1540.
46. Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005 May;435:207 EP –. Available from: http://dx.doi.org/10.1038/nature03459.
47. Bohorquez JC, Gourley S, Dixon AR, Spagat M, Johnson NF. Common ecology quantifies human insurgency. Nature. 2009 Dec;462:911 EP –. Available from: http://dx.doi.org/10.1038/nature08631.
48. Newcomb S. Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics. 1881;4:39–40.
49. Benford F. The Law of Anomalous Numbers. Proceedings of the American Philosophical Society. 1938;78(4):551–572. Available from: http://www.jstor.org/stable/984802.
50. Morzy M, Kajdanowicz T, Szymanski BK. Benford's Distribution in Complex Networks. Scientific Reports. 2016 Oct;6:34917 EP –. Available from: http://dx.doi.org/10.1038/srep34917.
51. Durtschi C, Hillison W, Pacini C. The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. J Forensic Account. 2004 01;5.
52. Jimenez R, Hidalgo M, Klimek P. Testing for voter rigging in small polling stations. Science Advances. 2017;3(6). Available from: http://advances.sciencemag.org/content/3/6/e1602363.
53. Dunn HL. Vital Statistics of the United states. Government of the United States of America; 1945.
54. OECD2018. Suicide rates (indicator), 10.1787/a82f3459-en;. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/suicide-rates/indicator/english_a82f3459-en.

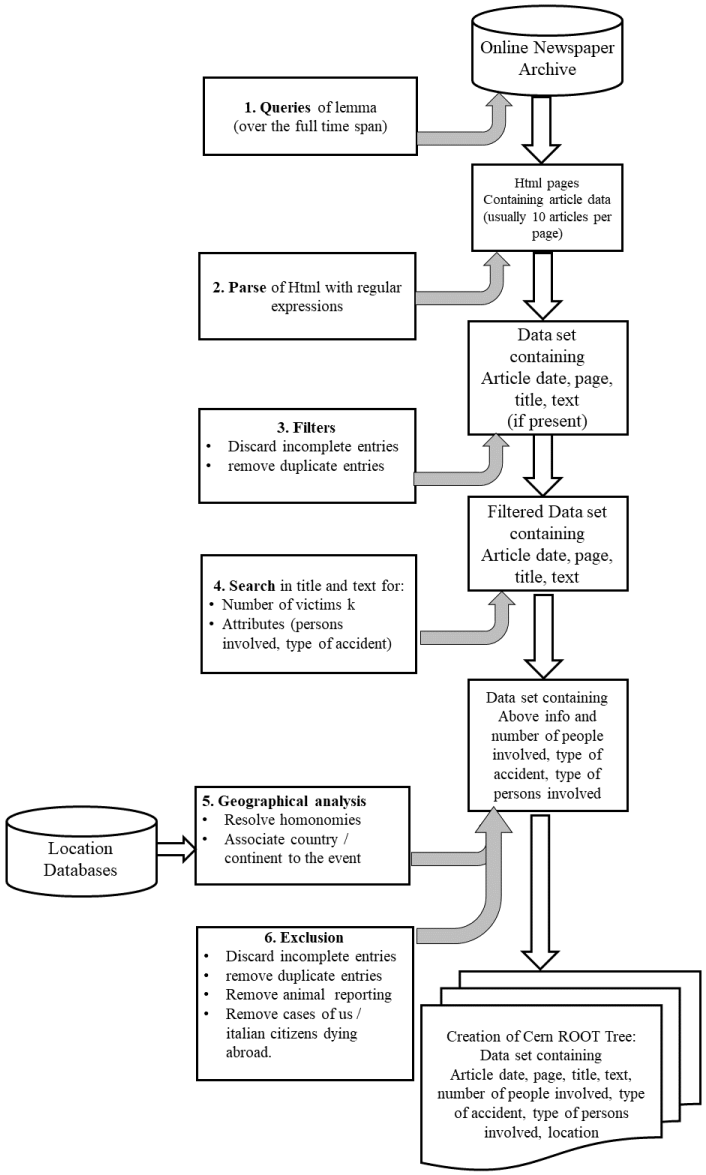

Figure 1 Scheme of the processing steps from the query on the newspaper archive to the creation of the database. Each query of a specific lemma on a newspaper results in a data set which is then subsequently analyzed.

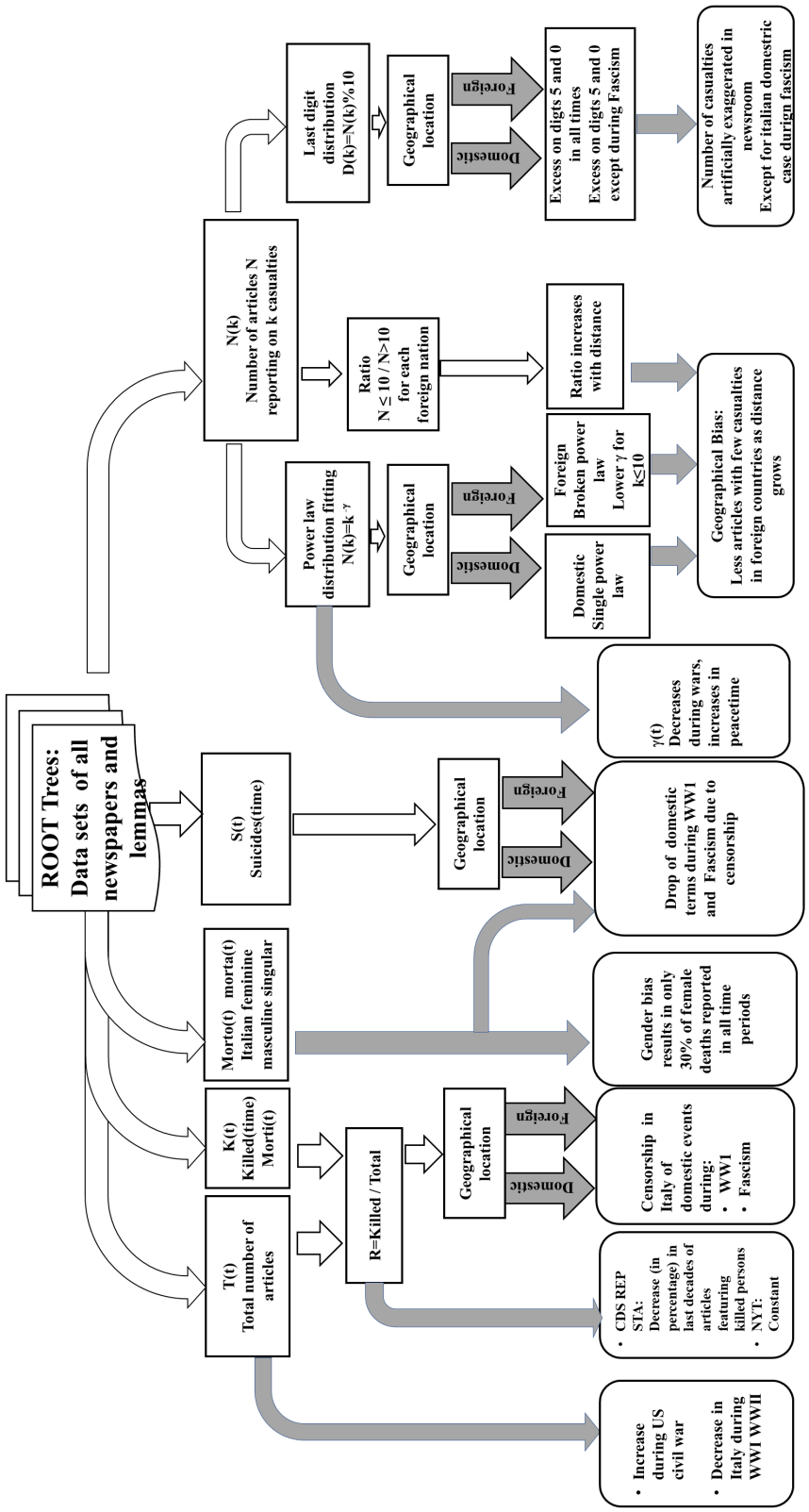

Figure 2 Scheme of the various algorithms employed on the databases and the information they provide.

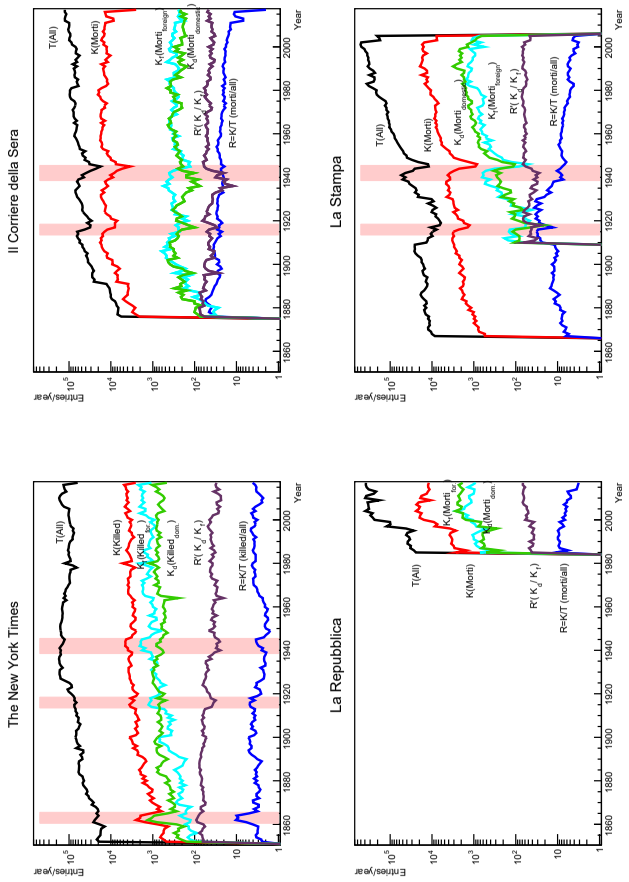

Figure 3 Time profile of datasets retrieved from the newspaper archives. Each panel shows: a) $T(t)$: Total number of articles present in the archive (black), b) $K(t)$: number of articles returned upon query of the word 'killed' or 'morti' (red), c) K_f: Number of foreign events (cyan), d) K_d: Number of domestic events (green), e) $R' = K_d/(K_f)$: Ratio domestic/foreign events $\times 10^2$ (purple), f) $R = K/T$: (killed/total or morti/total) $\times 10^2$ (blue). Red bands indicate the time of US Civil War (NYT), WW1 and WW2. It is possible to see the increase of T and K in NYT during Civil War and the corresponding decreases in STA and CDS during the two World Wars. For the same papers it is also evident a gradual decrease of K and K' during Fascist period (1923-1943) due to censorship (see text).

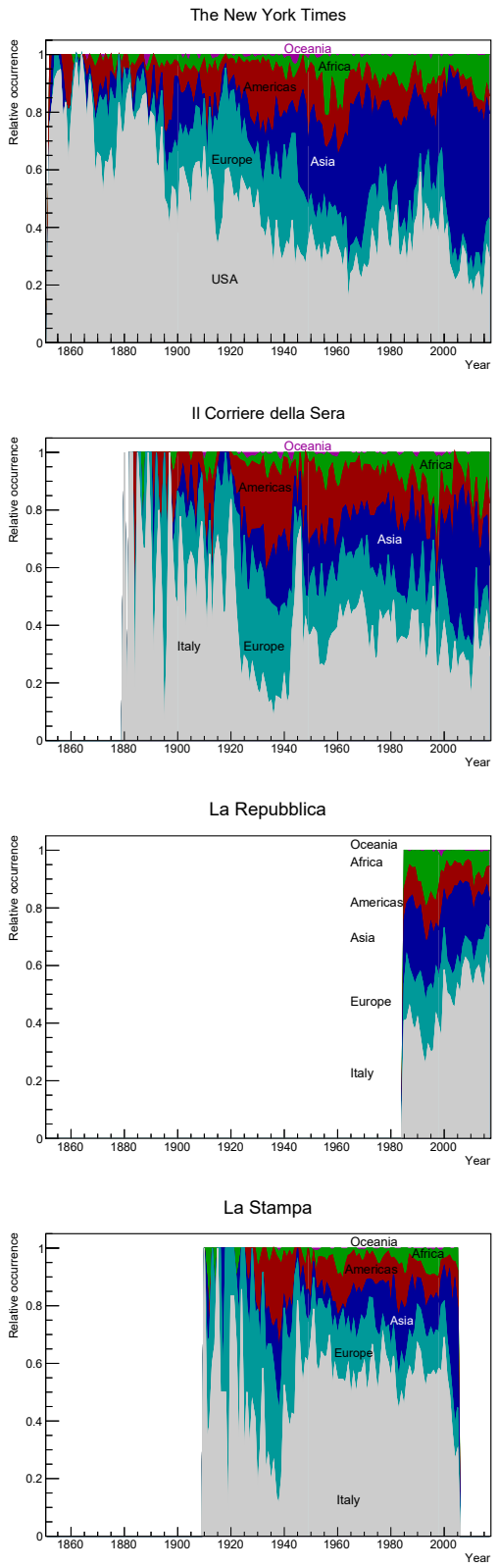

Figure 4 Normalized ratio of articles as function of time for domestic and foreign deaths in the various continents. It is possible to see how in the NYT the percentage of domestic events decreases with time. After WW2, the percentage of Asian events increases rapidly. In CDS and STA it is possible to see the decrease of domestic events during Fascism and the sharp peak after WW2.

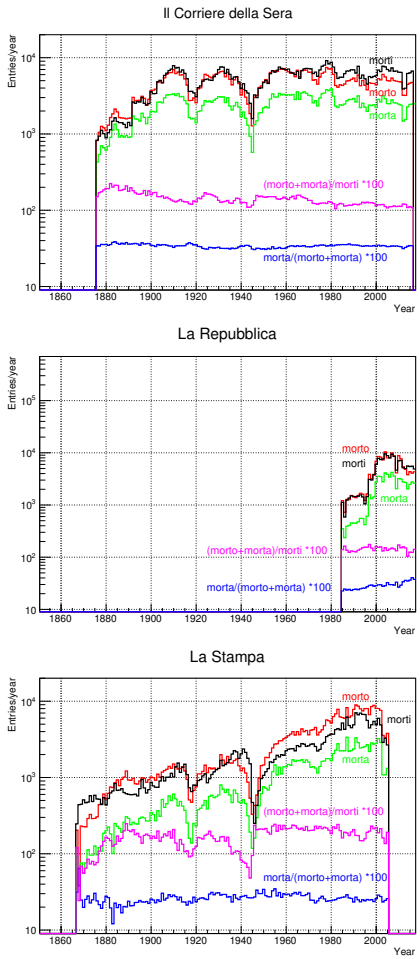

Figure 5 Number of occurrences for singular (*morto, morta*) vs plural (*morti*). Magenta is the ratio $(k = 1)/(k \geq 2)$, (*morto+morta)/morti*. The blue ratio refers to the percentage of female deaths, *morta/(morto+morta)*. Both ratios are $\times 10^2$). The percentage of $k = 1$ events halves for CDS over time and is roughly constant for STA after $\simeq 1890$, with drops during WW1 and Fascism. The reporting of female deaths is constant to $\simeq 30\%$ for CDS and STA, and doubles from 20% to 40% for REP.

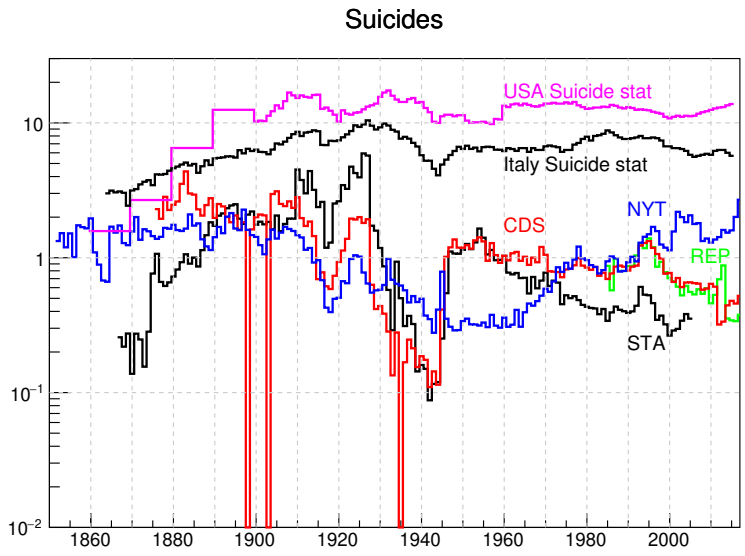

Figure 6 Percentage of articles containing the keyword *'suicide/suicidio'* for the four newspapers and the rate of suicides (cases per 100,000 population) vs time[20, 53, 54]. Note the sharp drops during the American Civil War (for NYT only), WW1 and WW2, both in the articles and in the rate. The drop during Fascist government is also visible. In 1898, 1902 and 1935 there are no articles retrieved with the keyword.

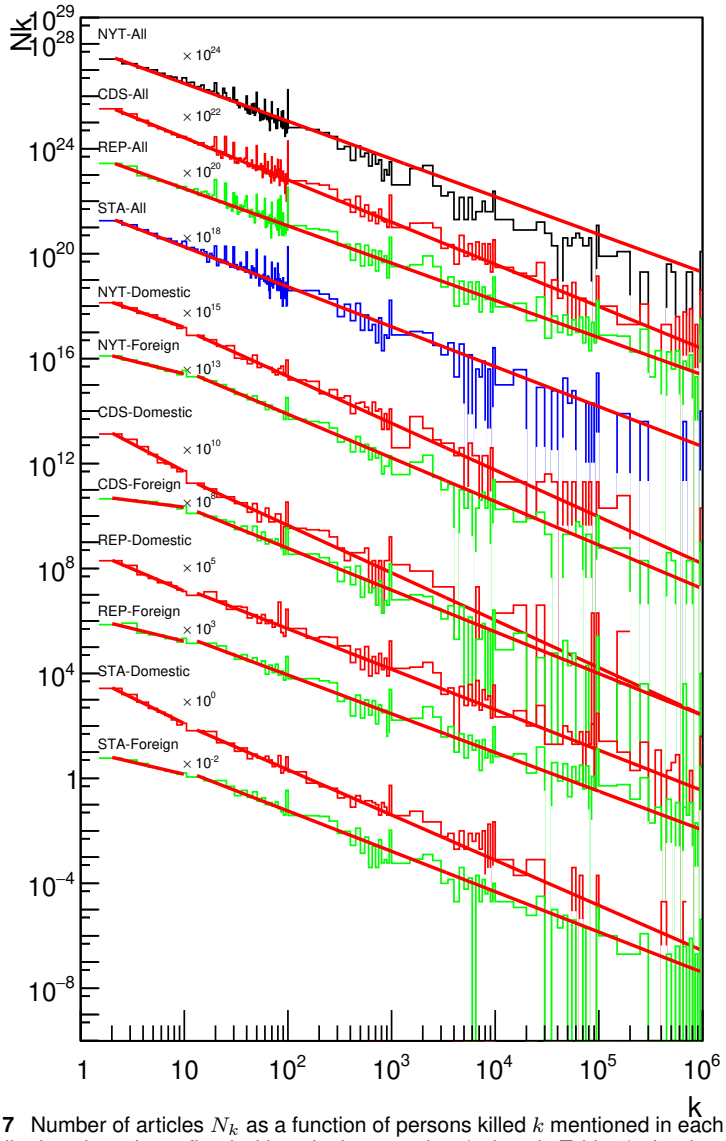

Figure 7 Number of articles N_k as a function of persons killed k mentioned in each article. The full distributions have been fitted with a single power law (values in Table 2), the domestic/foreign ones with two power laws γ_L $(2 \leq k \leq 10)$ and γ_H $(k > 10)$. For all foreign events there is a break in the spectral index $\gamma_L < \gamma_H$ due to missing events not being reported (from 91% for REP to 122% for CDS). In domestic events $\gamma_L > \gamma_H$ for Italian papers (over-reporting of low k events, from 47% in STA to 71% for CDS). In NYT whereas the decrease in γ_L for domestic events is lower (over-reporting of 4%) than for foreign ones (under-reporting of 98%), hinting to a higher degree of internationalization of this publication (Table 3).

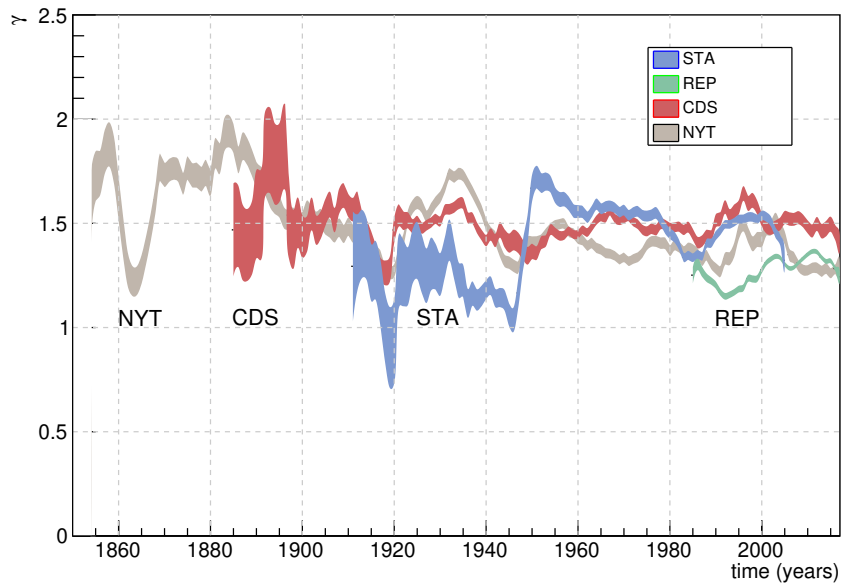

Figure 8 Spectral index γ as a function of time (running average on current and 4 preceding years). γ decreases during conflicts such as the Civil and the World Wars due to the high k events. The bands indicate the error from the fit.

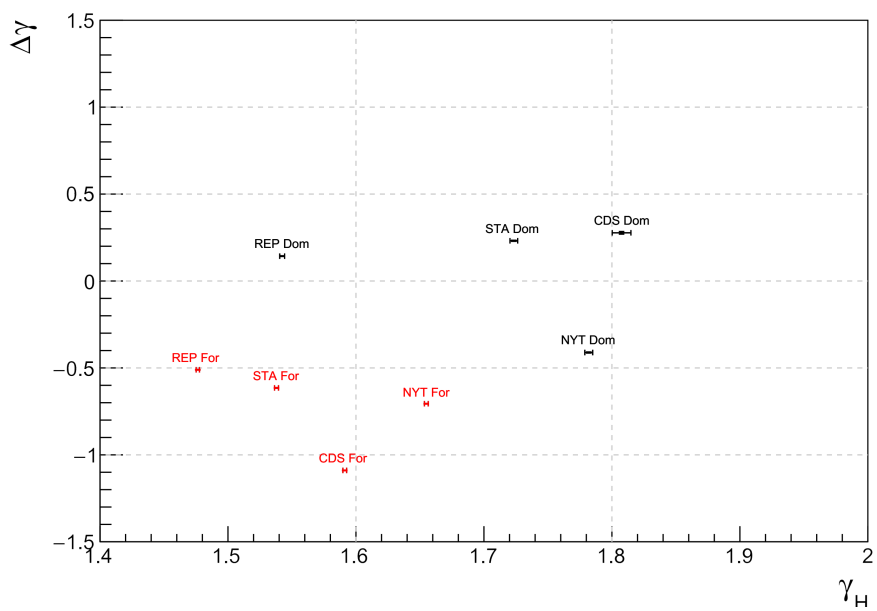

Figure 9 $\Delta\gamma = \gamma_L - \gamma_H$ vs γ_H obtained from the fit of power-law for the distributions of domestic and foreign events. $\Delta\gamma \simeq 0$ means a pure power law over the whole distribution whereas a negative (positive) $\Delta\gamma$ implies a lack (excess) of articles with a small $(2 \leq k \leq 10)$ number of deaths. An higher (lower) value of γ_H implies more emphasis on low (high) values of k. In all newspapers, foreign distributions have a lower $\Delta\gamma$ than domestic ones and a lower value of γ_H, showing that high k events have an higher importance over the low k ones. NYT has the smallest differences in $\Delta\gamma$, a sign of greater uniformity of treatment between domestic and foreign events.

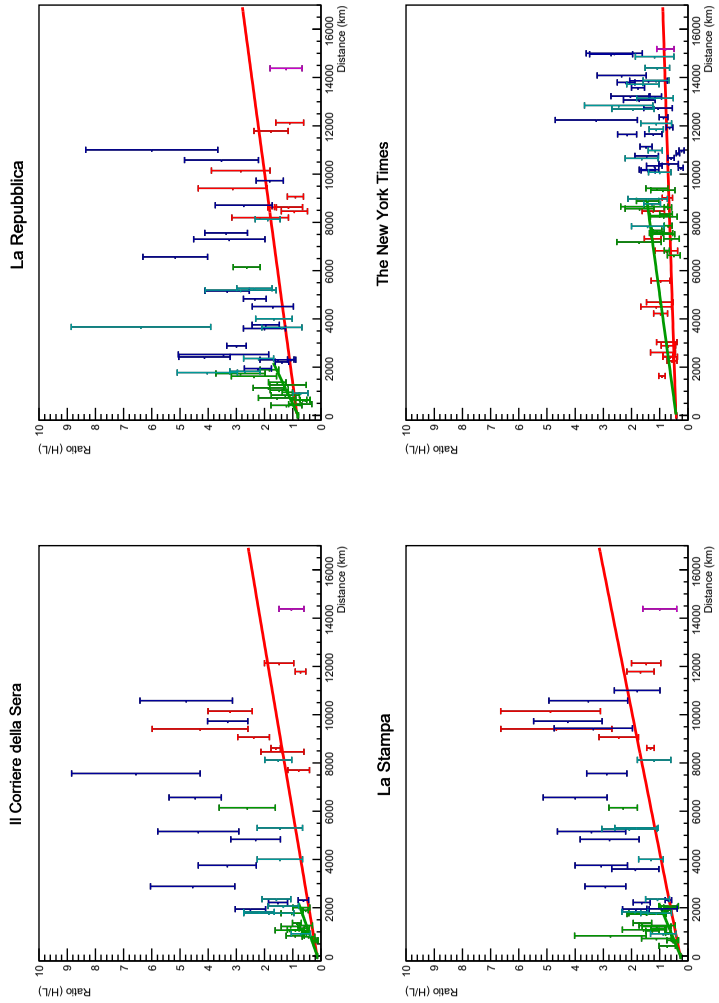

Figure 10 Distribution of the ratio $W_i = N_H/N_L$ vs distance between Rome (for CDS, REP, STA) and Washington (for NYT) and the capital of world nation i with more than 30 events in each newspaper dataset. Color denotes the continents: Green - Europe, Bue: Africa, Cyan: Asia, Red: America, Purple: Oceania. Red lines indicate linear fit on the whole world, Green lines indicate linear fit only on Europe/USA.

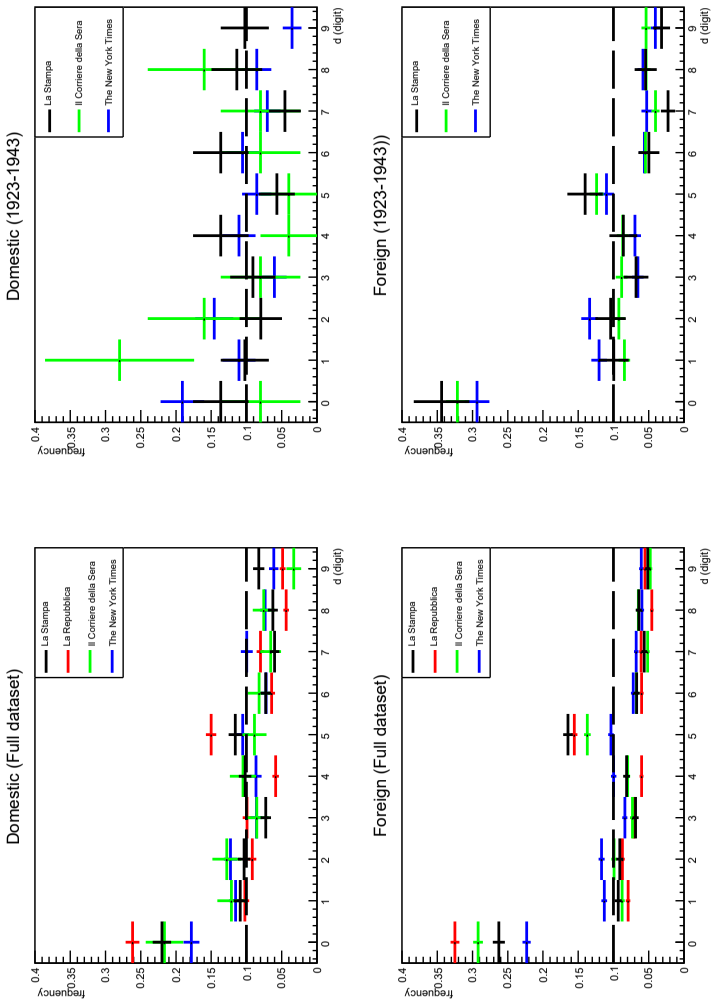

Figure 11 Relative occurrence of least significant digit for $10 < k \leq 100$. Top: domestic, Bottom: foreign events. Left: full dataset, Right: period of Fascist government in Italy. The dashed line shows the expected value from a random distribution (0.1). The values of '1' and '2' are closer to this value, with some defect in Italian newspapers and excess in NYT. '3' and '4' are below expectations and their defect is close to the excess of '5' events. The least probable digits are '6' to '9', usually rounded in excess to '0'. Only domestic events during Fascist government in Italy are not incompatible with equiprobable distribution.

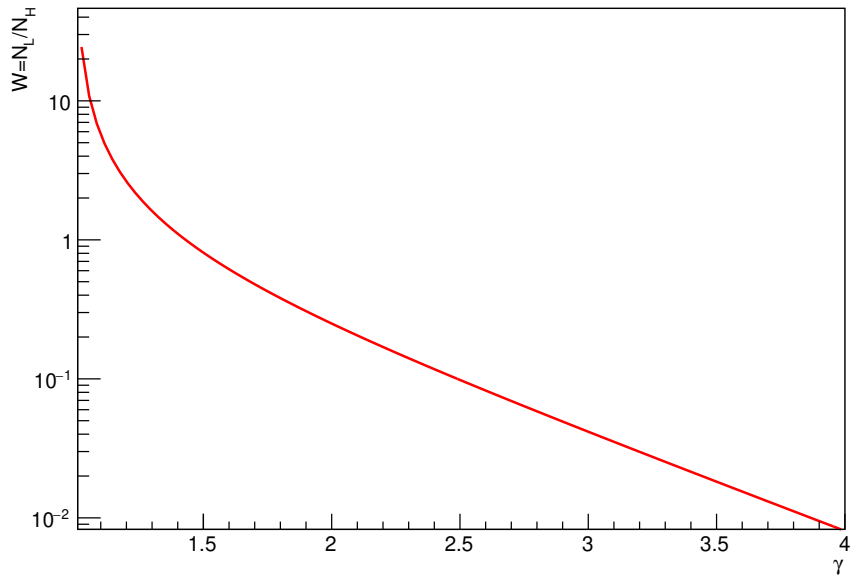

Figure 12 Ratio $W = N_H/N_L$ as a function of γ for a single power-law distribution of N_k. A value of $\gamma = 1.43$ results in $W = 1$, an equal number of articles with $2 \leq k \leq 10$ and $k > 10$.

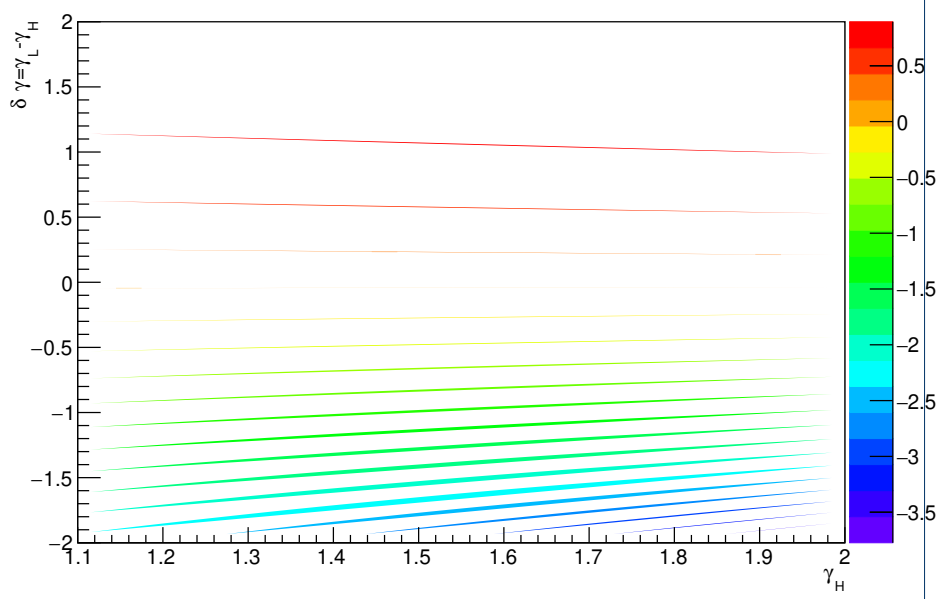

Figure 13 M, amount of lack ($M < 0$) or excess ($M > 0$) articles in respect to a pure power law distribution ($\delta_\gamma = 0$).

Figures

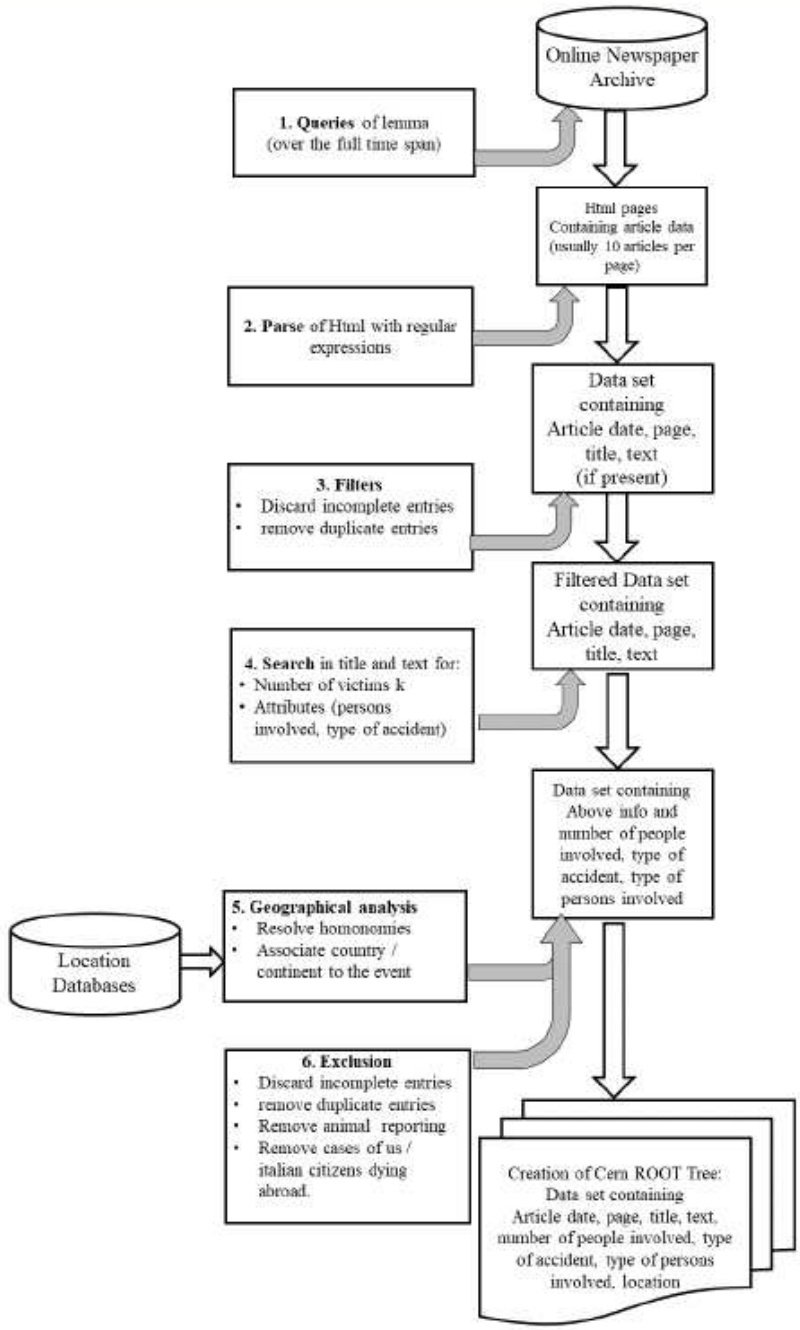

Figure 1 Scheme of the processing steps from the query on the newspaper archive to the creation of the database. Each query of a specific lemma on a newspaper results in a data set which is then subsequently analyzed.

Figure 1

Scheme of the processing steps from the query on the newspaper archive to the creation of the database. Each query of a specific lemma on a newspaper results in a data set which is then subsequently analyzed.

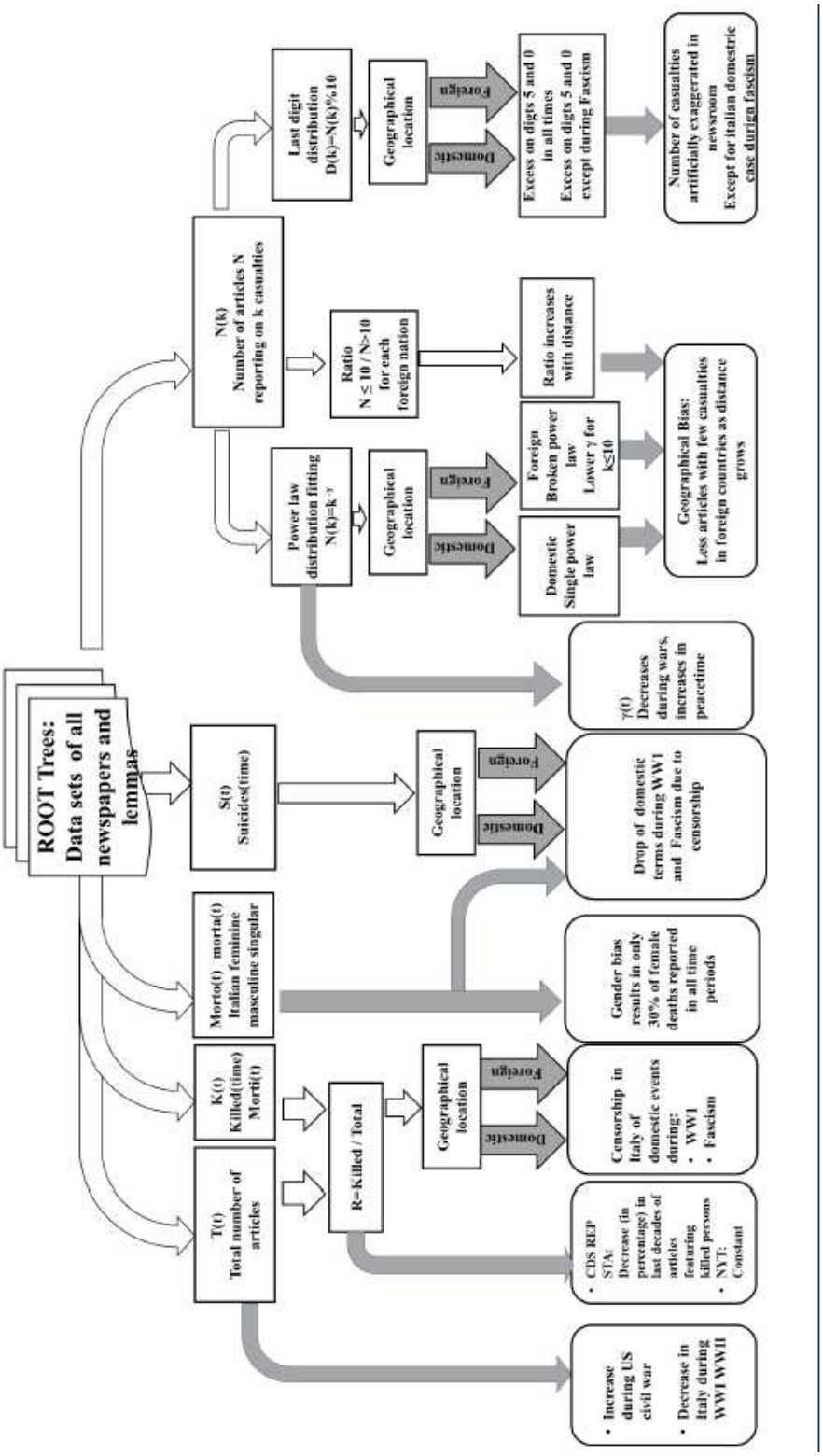

Figure 2

Scheme of the various algorithms employed on the databases and the information they provide.

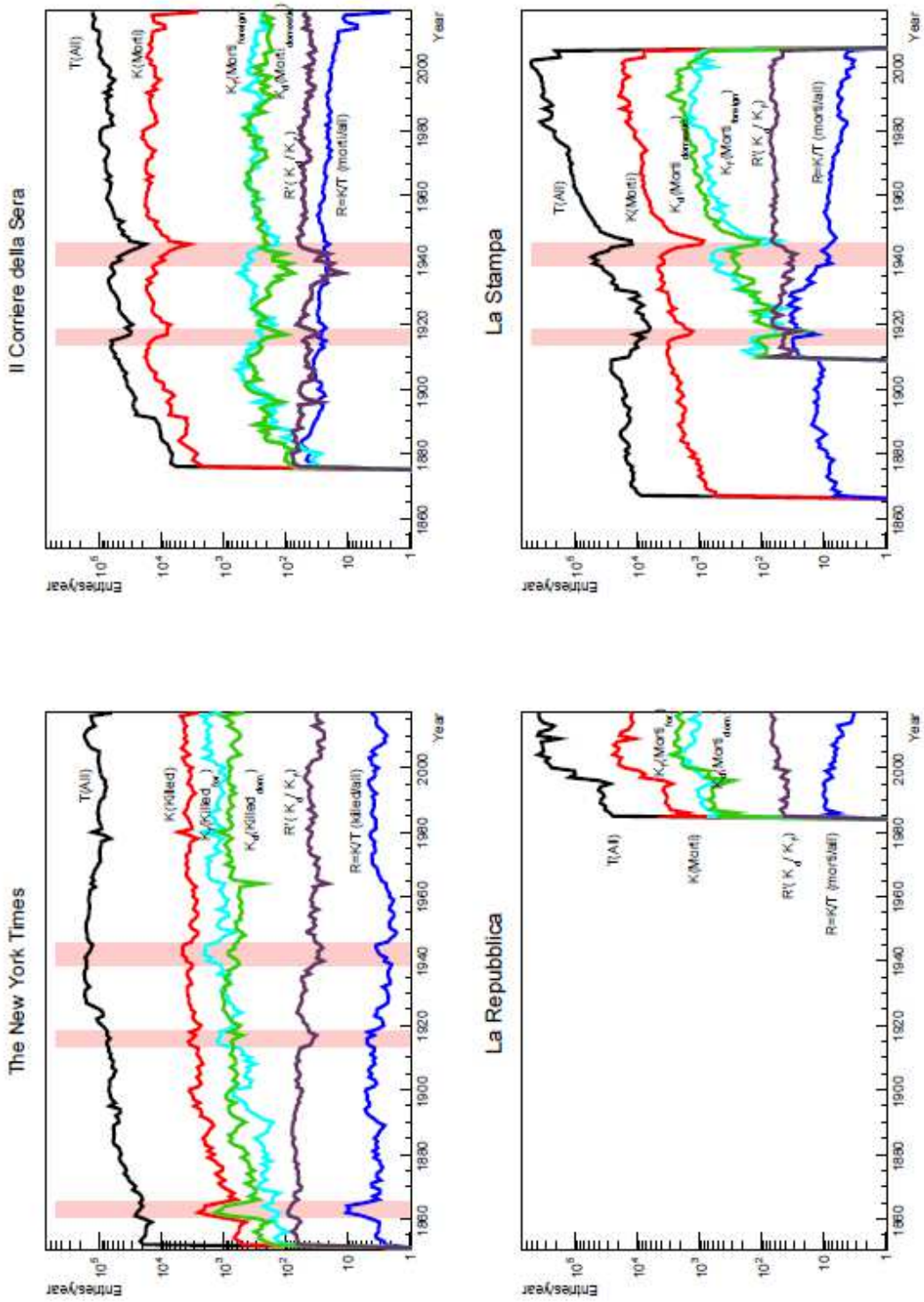

Figure 3

Time profile of datasets retrieved from the newspaper archives. Each panel shows: a) T(t): Total number of articles present in the archive (black), b) K(t): number of articles returned upon query of the word ‘killed’ or ‘morti’ (red), c) Kf : Number of foreign events (cyan), d) Kd: Number of domestic events (green), e) R0 = Kd=(Kf): Ratio domestic/foreign events x10^2 (purple), f) R = K=T : (killed/total or morti/total) x10^2 (blue). Red bands indicate the time of US Civil War (NYT), WW1 and WW2. It is possible to see the

increase of T and K in NYT during Civil War and the corresponding decreases in STA and CDS during the two World Wars. For the same papers it is also evident a gradual decrease of K and K0 during Fascist period (1923-1943) due to censorship (see text).

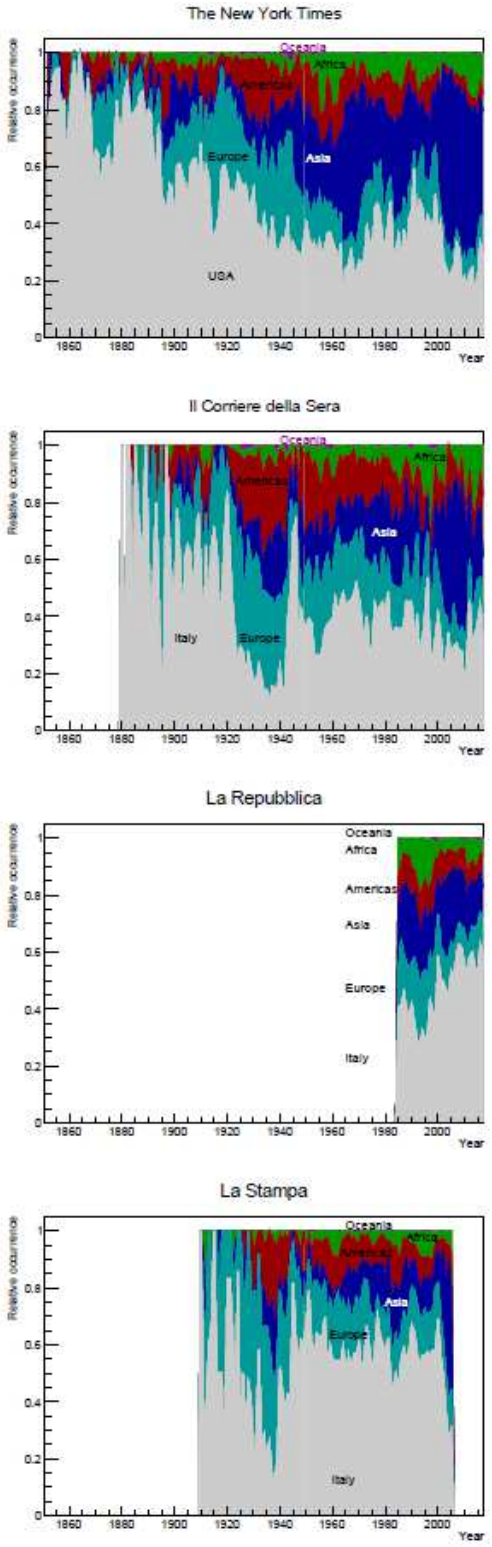

Figure 4

Normalized ratio of articles as function of time for domestic and foreign deaths in the various continents. It is possible to see how in the NYT the percentage of domestic events decreases with time. After WW2,

the percentage of Asian events increases rapidly. In CDS and STA it is possible to see the decrease of domestic events during Fascism and the sharp peak after WW2.

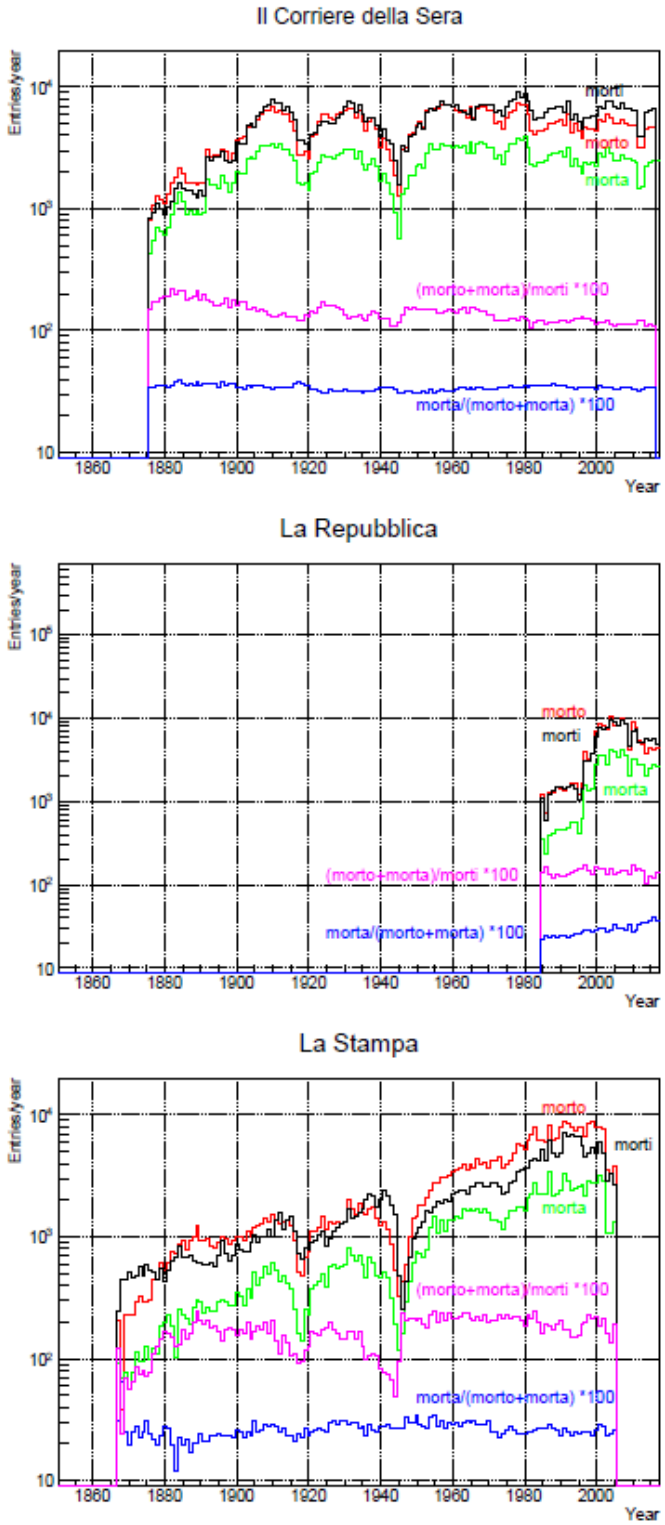

Figure 5

Number of occurrences for singular (morto, morta) vs plural (morti). Magenta is the ratio (k = 1)=(k ≥ 2), (morto+morta)/morti. The blue ratio refers to the percentage of female deaths, morta/(morto+morta). Both ratios are 102). The percentage of k = 1 events halves for CDS over time and is roughly constant

for STA after ' 1890, with drops during WW1 and Fascism. The reporting of female deaths is constant to ' 30% for CDS and STA, and doubles from 20% to 40% for REP.

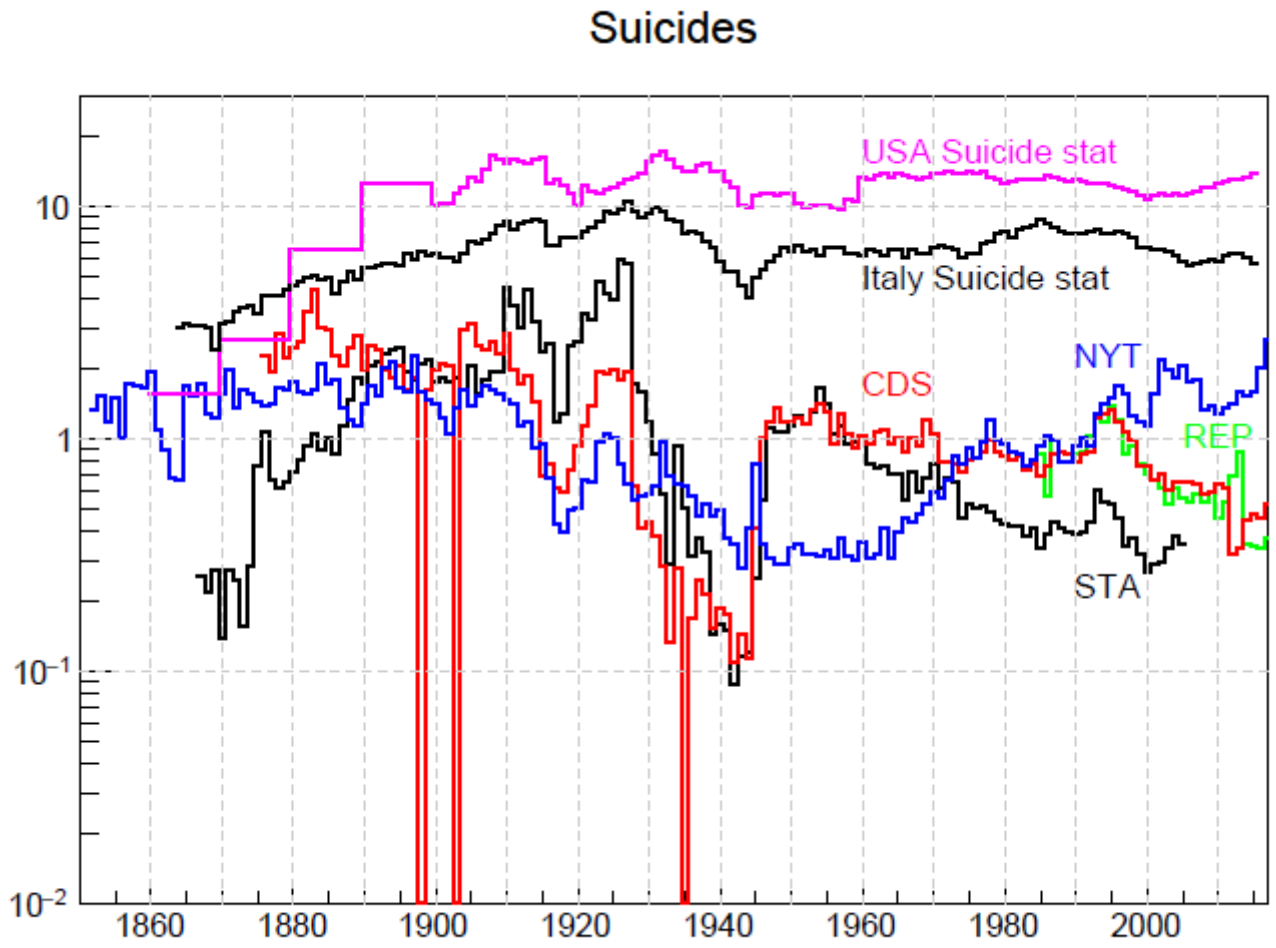

Figure 6

Percentage of articles containing the keyword 'suicide/suicidio' for the four newspapers and the rate of suicides (cases per 100,000 population) vs time[20, 53, 54]. Note the sharp drops during the American Civil War (for NYT only), WW1 and WW2, both in the articles and in the rate. The drop during Fascist government is also visible. In 1898, 1902 and 1935 there are no articles retrieved with the keyword.

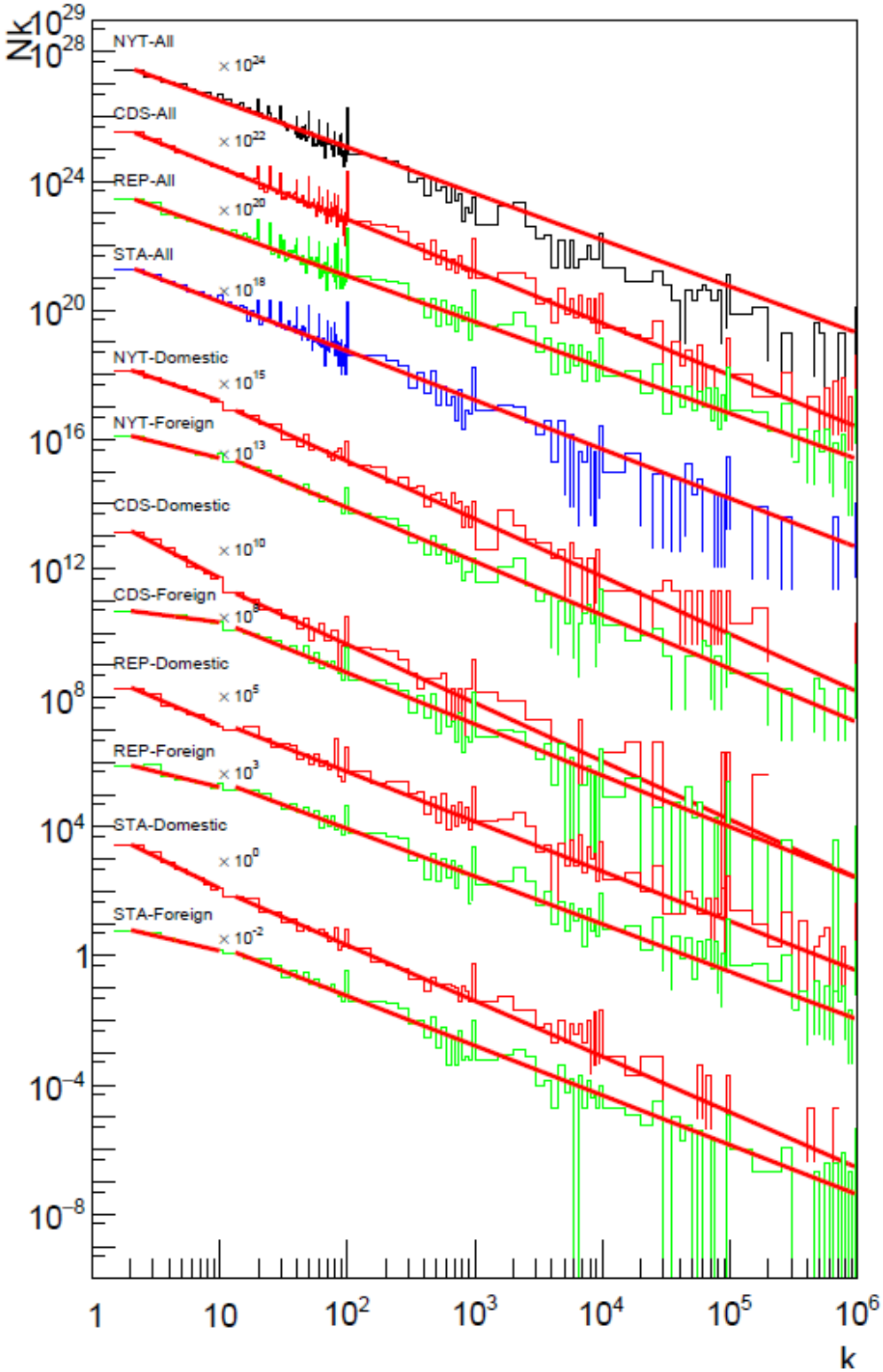

Figure 7

Number of articles Nk as a function of persons killed k mentioned in each article. The full distributions have been fitted with a single power law (values in Table 2), the domestic/foreign ones with two power laws L ($2 \leq k \leq 10$) and H ($k > 10$). For all foreign events there is a break in the spectral index L < H due to missing events not being reported (from 91% for REP to 122% for CDS). In domestic events L > H for Italian papers (over-reporting of low k events, from 47% in STA to 71% for CDS). In NYT whereas the

decrease in L for domestic events is lower (over-reporting of 4%) than for foreign ones (under-reporting of 98%), hinting to a higher degree of internationalization of this publication (Table 3).

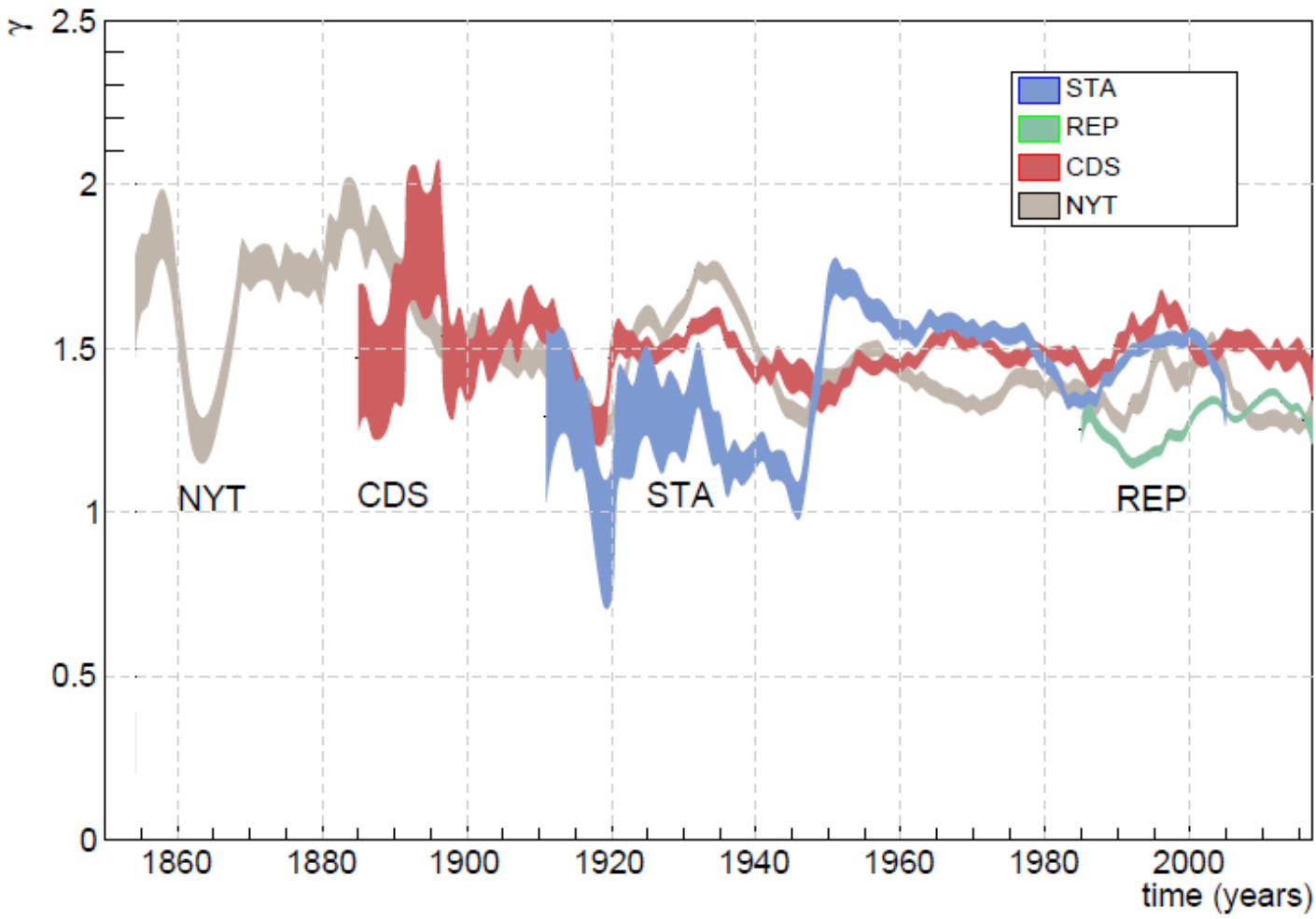

Figure 8

Spectral index as a function of time (running average on current and 4 preceding years). decreases during conflicts such as the Civil and the World Wars due to the high k events. The bands indicate the error from the fit.

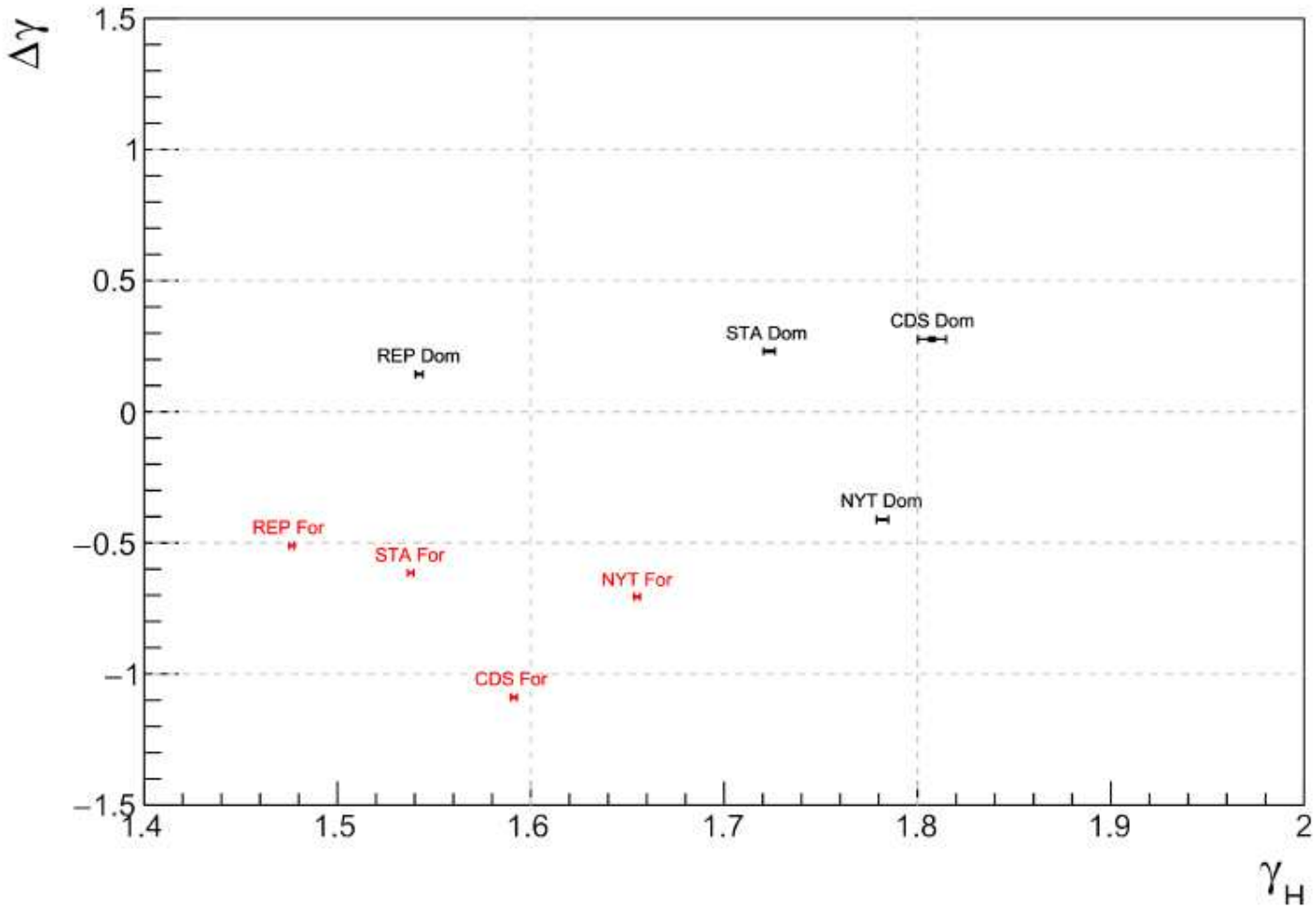

Figure 9

y = yL - yH vs yh obtained from the fit of power-law for the distributions of domestic and foreign events. y 0 means a pure power law over the whole distribution whereas a negative (positive) y implies a lack (excess) of articles with a small ($2 \leq k \leq 10$) number of deaths. An higher (lower) value of H implies more emphasis on low (high) values of k. In all newspapers, foreign distributions have a lower y than domestic ones and a lower value of H, showing that high k events have an higher importance over the low k ones. NYT has the smallest differences in y , a sign of greater uniformity of treatment between domestic and foreign events.

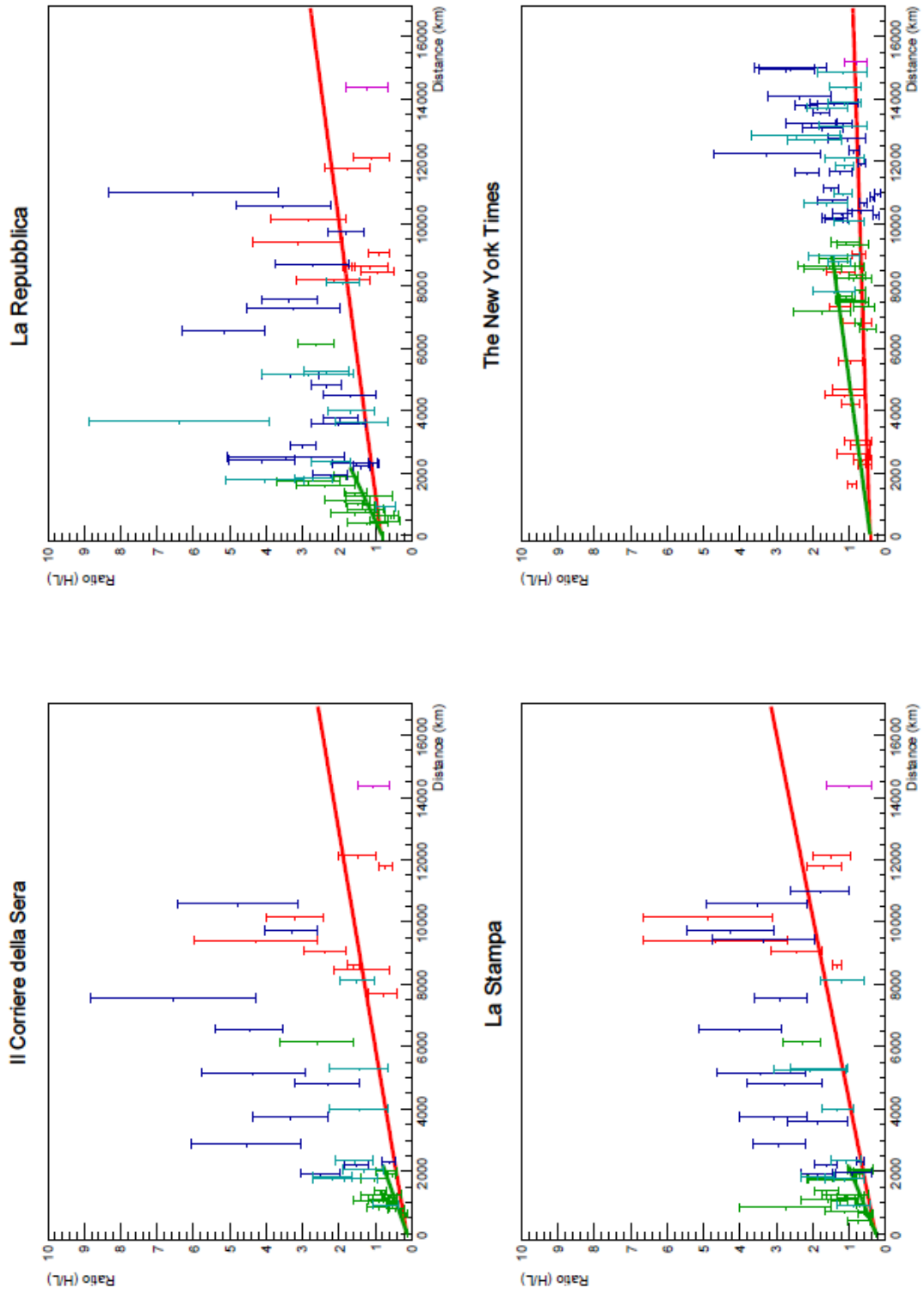

Figure 10

Distribution of the ratio Wi = NH=NL vs distance between Rome (for CDS, REP, STA) and Washington (for NYT) and the capital of world nation i with more than 30 events in each newspaper dataset. Color denotes the continents: Green - Europe, Bue: Africa, Cyan: Asia, Red: America, Purple: Oceania. Red lines indicate linear fit on the whole world, Green lines indicate linear fit only on Europe/USA.

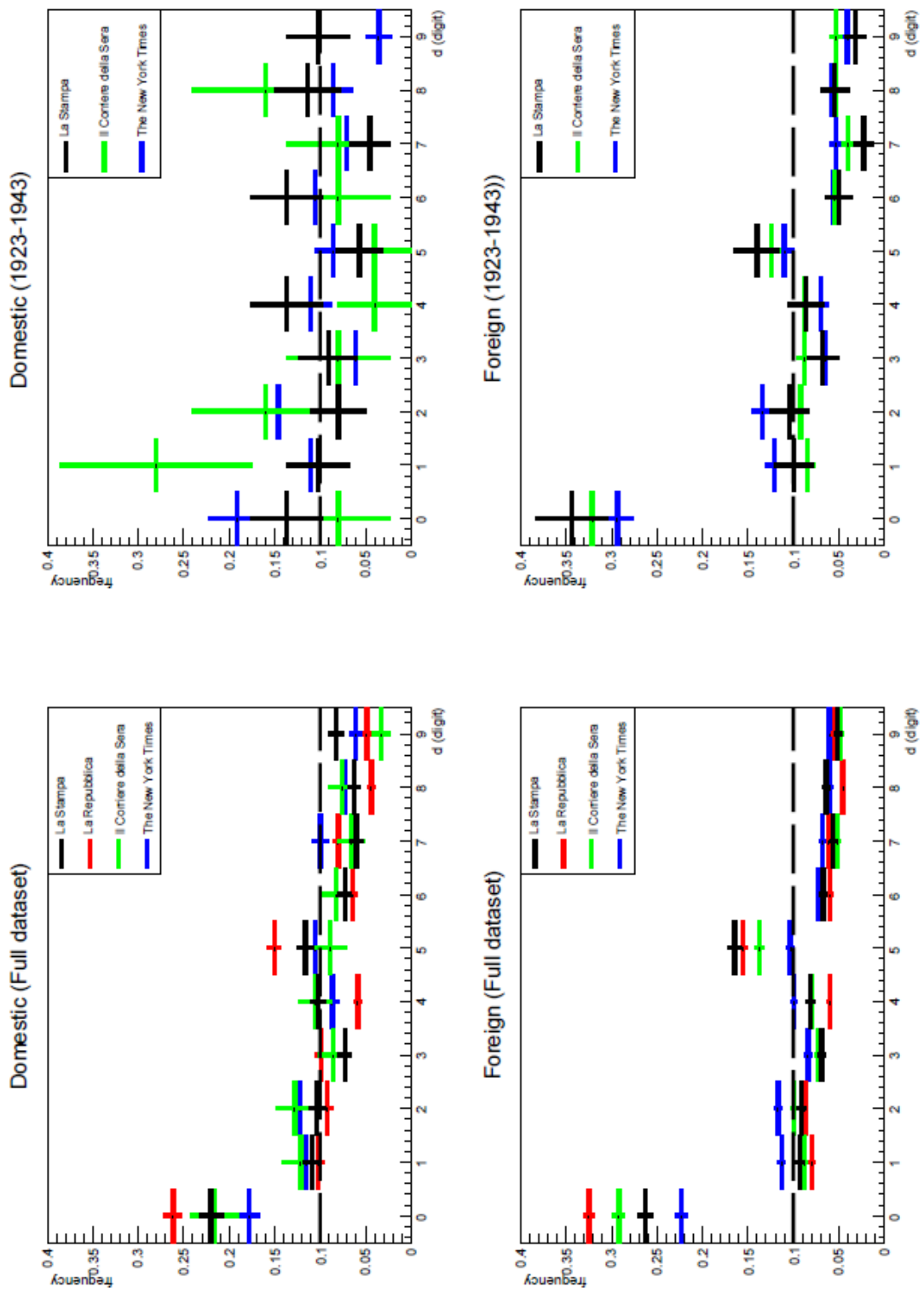

Figure 11

Relative occurrence of least significant digit for 10 < k ≥ 100. Top: domestic, Bottom: foreign events. Left: full dataset, Right: period of Fascist government in Italy. The dashed line shows the expected value from a random distribution (0.1). The values of '1' and '2' are closer to this value, with some defect in Italian newspapers and excess in NYT. '3' and '4' are below expectations and their defect is close to the excess

of ‘5’ events. The least probable digits are ‘6’ to ‘9’, usually rounded in excess to ‘0’. Only domestic events during Fascist government in Italy are not incompatible with equiprobable distribution.

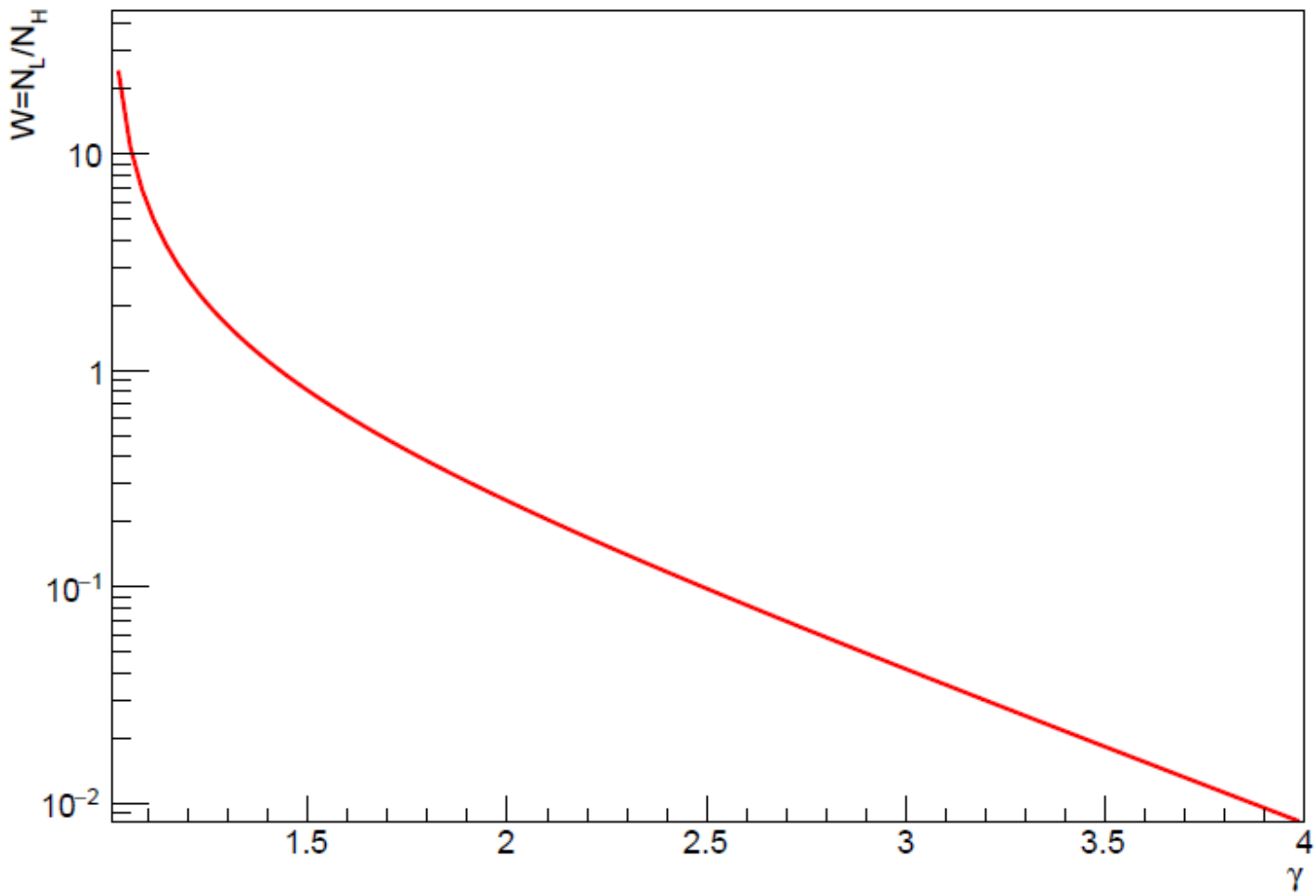

Figure 12

Ratio W = NH=NL as a function of for a single power-law distribution of Nk. A value of = 1:43 results in W = 1, an equal number of articles with 2≤ k ≤ 10 and k > 10.

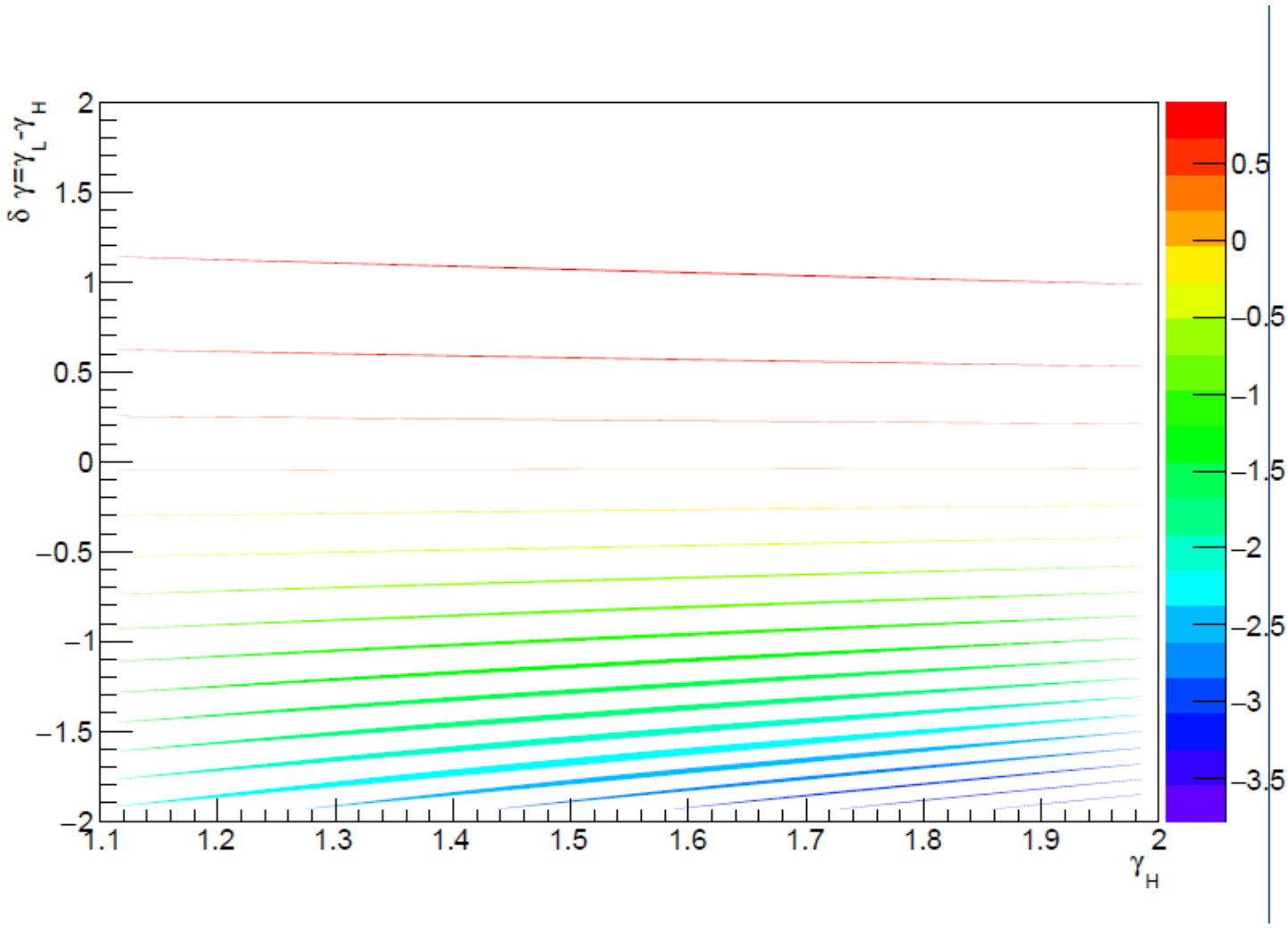

Figure 13

M, amount of lack (M < 0) or excess (M > 0) articles in respect to a pure power law distribution (Sy = 0).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- violentdeathsupp.pdf