A rank-size approach to the analysis of socio-economics data

University of Macerata

Department of Economics and Law

Quantitative Methods for Economics Policy

cycle: XXX

Ph.D. Candidate: Dr. Valerio Ficcadenti
Supervisor: Prof. Roy Cerqueti
Coordinator: Prof. Maurizio Ciaschini

2018

Dedication

Ludovica, Family and Friends...

Acknowledgements

Prof. Roy Cerqueti for introducing me to simplicity of complexity.
Prof. Gurjeet Dhesi for his out of the box point of view.
Prof. Marcel Ausloos for his competence and realism.
Ing. Marco Ortenzi for his effective coding suggestions.

Contents

Introduction

A deep knowledge of social and natural phenomena is a necessary step for the comprehension of the world we live in. Related analyses can be performed under several perspectives.

From a pure, ideal point of view, there is the need of models which can be tailored on the reality in such a way that it is possible to have a broad and complete vision of the investigated phenomena. The achievement of this objective can be reached thanks to some theoretical models developed and treated with mathematical instruments. However, theoretical models can be conceptualized starting from empirical studies necessary for the assessment of the main properties of the considered problems.

This PhD thesis deals with a deep exploration of two key natural and human facts. The first one is related to the earthquakes, while the second one is associated to the content of the official speeches of the US Presidents. In particular, our aim is to define the extent of the economic damages deriving from earthquakes on the basis of a large series of magnitudes over a rather large period. Furthermore, we investigate the economic impact of the speeches of the US Presidents on the financial market, with a specific reference to the Standard and Poor's 500. Our main objective is to give a contribution into the economic policy field by taking into consideration these phenomena from a different and innovative perspective. At the best of our knowledge, in these fields, the following approach is the first time it appears.

We employ several methodological tools. However, we can identify the ground of the analysis with the econophysic instruments related to the rank-size law. It is a set of different functions applied with the aim of exploring the properties of large sets of data, even when they are distributed over time and error bars are not clear because of peculiar sampling conditions.

[146, 147] originally introduced the power law and Pareto distribution with a unitary coefficient to explore the rank-size relationship in the field of

linguistics.
After that Zipf has presented the power law, many researchers have contributed into the field by analyzing the law properties and by applying variation of it to different scientific sections. In the following rows we cite some remarkable cases: [16, 143, 23] for the investigation into business size field; [67] for a biology approach; [49] case of fraud detection contribution; [53, 41, 30, 27] for the contribution in the context of economics geography; [66, 70] in analysing computer science dataset; [22] in the gaming field; [71, 144], in the context of music. In [111] a complete review of the rank-size approach is proposed. It is possible to list many other remarkable examples but here it seems enough for having in mind the breadth of possible applications .
The rank-size law presents some weakness in peculiar cases. For example the fits are slightly worse when the tails of the distribution are rich of outliers. Some example of these situations are reported in [116, 109, 52, 77]. These minor elements can be traced back to the difficulties in founding theoretical justification for the strong statistical regularity often represented (see [39, 136]). But, this point is also considered a further incentive for proceeding with the methodological research, and construct more general laws.

There is a high number of relevant studies that concerns the application of rank-size methods to text analysis; they constitute an early stage of the text mining manipulation or they are part of the Natural Language Processing field, see e.g. [98, 110, 8, 9, 50, 15].

In [38] there is a clarification of main critics about the studies of texts regularities with a frequentist approach. Indeed, the usage of Zipf's law for text analysis has been questioned by many researchers but it is still used.

During the years, many scholars have been modifying the law originally proposed by introducing extensions as in [113, 36]. Even Mandelbrot, with the Zipf-Mandelbrot's law (ZML, hereafter) has contributed to the evolution of the Zipf's law with two landmark studies [72, 73]. In [63] the Lavalette law (LL hereafter), which has a noticeable fit of rank-size relations even when Zipf's law fails to do it, is presented (see e.g.[27, 13]).

As proven by this literature overview, the rank-size relationship has been explored for several sets of data and the features of the family of functions that this name refers to can be considered strong enough to justify an econophysics approach to the data analysis.

In the chapter of this thesis that concern the earthquakes as well as those referring to the speeches of the USA presidents, we show and comment the results of the Levenberg-Marquardt Non-linear Least-Squares Algorithm (see [65, 69, 76]). Indeed, we adopt this method to implement many best fit procedures on the data for different types of rank-size functions. The resulting estimations are used for grasping relevant economics conclusion from different perspective.

The Chapter 1 is devoted to explore the features of the magnitude of the earthquakes occurring in italy between January 24th, 2016 and January 24th, 2017 in order to elaborate a proposal of cost indicators. The well known tragic seismic events with epicenters at Accumuli, Visso, Ussita, Castelsantangelo sul Nera, Norcia and Montereale are occurred during the considered span of time. On this dataset we develop two different rank-size analysis: the standard ZML and the Universal Law (UL hereafter) proposed by [13].
The idea of designing an economic impact measure of the earthquakes is based on the evident fact that exists a cause-effect relationship for the magnitude of earthquakes and the economic cost deriving from them. We draw attention to the role of the infrastructure resistance into the relationship between the damages and the seismic events sequences, so we conjecture different form of cost indicators.

In Chapter 2, we construct and analyze one of the biggest dataset of the US presidential speeches, with the aim of exploring their rhetoric regularities. Under our perspective, the speeches are considered as complex systems. The words of each speech are sorted in accord to their frequencies and analyzed through rank-size laws.
The dataset building phase is a pillar of this study. We apply a web scraping routine on the Miller Center web site in order to obtain a list of 978 raw of transcripts of the speeches. With a text refining process we reduce them to 951 and for each one a tokenization phase is employed to divide and store each single word. The ZML is fitted to each speech individually.
Thanks to the described process it is possible to collect interesting information on all the 45 United States Presidents and on the USA history, from April 30th, 1789 to February 28th, 2017. The results show some remarkable regularities inside each speech and among them.

In Chapter 3 we study a sub-selection of the dataset employed in Chapter 2. In particular we deal with a quantitative analysis of the informative content of the words pronounced only once in each speech by US Presidents, the so-called hapaxes. We search for the relevance of such rare words and so we implement a rank-size procedure of Zipf-Mandelbrot type for discussing the hapaxes' frequencies regularity over the overall set of speeches that covers the entire US history. Starting from the obtained rank-size law, we detect the core of the hapaxes set. We further show that this core of hapaxes itself can be well fitted through a Zipf-Mandelbrot law and that contains elements producing deviations at the low ranks between scatter plots and fitted curve – the so-called king and vice-roy effect. Some rhetoric and semantic framework insights are derived from the obtained findings about the US Presidents messages.

In Chapter 4, we explore the economic content of the speeches of the US presidents. Furthermore we collect a list of economical terms and we quantify their presence into each transcript in order to evaluate the impact of such a bunch of words on some economical and financial indexes.

Rank-size analysis intervenes in this context through Kendall τ correlation. In particular, we explore the rank-correlation between the normalized absolute and relative frequencies of the economic terms in the speeches and the normalized returns, closing prices and volumes of the Standard & Poor's 500 series. Three paradigmatic cases are considered: the realization of the considered index in the same days of the speeches, one day ahead and one day back. In so doing, we include in the analysis the possibility of an anticipatory role of the economics speeches, as well as the direct impact or the consequence of market realizations. As a side scientific product, we also analyze the distances between the series of frequencies of the economic terms and the Standard & Poor's 500 index. At this aim, we adopt a probabilistic approach but also a mere topological perspective. In fact, entropy measures and several concepts of vectoral distances are compared.

Chapter 1

The economic cost indicator of earthquakes. The case of the Italian seismic events of 2016-2017

Despite of the fact that Italy comes from an eventful seismic history, the legislator has taken a series of myopic decisions, leaving to Italy a dramatic exposure to damages generated by earthquakes. In fact, Italy has an old and poorly maintained housing system that is very sensitive to seismic events, especially when they are in series and above a certain threshold.
The seismologists classify the entire planet by using different criteria based on the probability of occurring an earthquake into specific zones. Differently from the Italian case, Japan with its even higher risk position of having a natural disaster, has gained a very good experience in anti-seismic buildings development and it has developed an effective strategy for reducing the damages and so the social cost.

In this chapter we take in consideration the Italian seismic events occurred between January 24^{th}, 2016 and January 24^{th}, 2017; we include the cases of August, 24^{th} 2016 in Accumuli and October, 26^{th} and 30^{th} 2016 in Ussita and Norcia beyond that minor shocks happened in Italy within the considered span of time. The total number of temblors observed in the considered span of 365 days is 978, within a Richter magnitude range: [3.1 - 6.5]. The maximum level corresponds to the earthquakes of October 30^{th}, 2016 with municipality of Norcia as epicentrum. While, the minimum considered level

3.1 comes from different considerations. In the raw data downloaded from the Italian National Institute of Geophysics and Vulcanology (INGV) there is an enormous amount of minor shocks registered that is not able to generate damages. Indeed, in according to the United State geological survey, the earthquakes with magnitude inferior to 3.1 has law chances of creating observable problems. Furthermore, adopting a cutoff at 3.1, the catalogue incompleteness problem is avoided. The issue is met because the analyzed period is a very peculiar one with many relevant shocks occurred in a short span of time. It is such special that SISMIKO, the emergency seismic network at INGV, has installed an adding detection system in order to support the permanent one around the original epicentral area. It is required in order to lower the risk of not capturing the aftershocks that could remain out from the catalogue. This kind of problem requires an adding effort from INGV which has to manipulate the raw data for completing the dataset (see [99]). On the catalog completeness problems there are several studies recently developed. For example in [75] the authors have determined a threshold of $M_c = 2.7$ for the revised catalog of shocks happened after the so called Amatrice earthquake (it is the shock occurred the August 24^{th},2016, known with the name of the town where the caused damages were the most). They wrote that the lower bound that ensures the completeness of the dataset could rise to 3.1 and this is concordant with [28].

In Figure 1.1 we can easily detect the highest picks occurred after August, 24^{th} 2016. Consequently one can assert that the majority of the observations in the dataset should not be affected by the incompleteness problem; therefore we can consider an $M_c = 2.5$ for the time series from January, 24^{th} 2016, to August, 24^{th} 2016 in accordance with [115, 120]. Anyway, the most prudential restriction at 3.1 is considered and it makes the dataset adequate for the cost analysis of the earthquakes.
The complete version of the catalog has very interesting properties from a rank-size point of view, especially when the UL is employed; but it cannot be fruitfully used for evaluating economic cost as it is shown in the next sections.

In this chapter we use two different rank-size laws to elaborate some functional relationships that allow us to provide indicators of the damages caused by different shocks. We implement many best fit procedures on the ZML and on the UL that is an extension of the LL mainly differentiated by its 5 parameters. In order to do so we take up the Levenberg-Marquardt algorithm with restriction on parameters that have to be positive. This

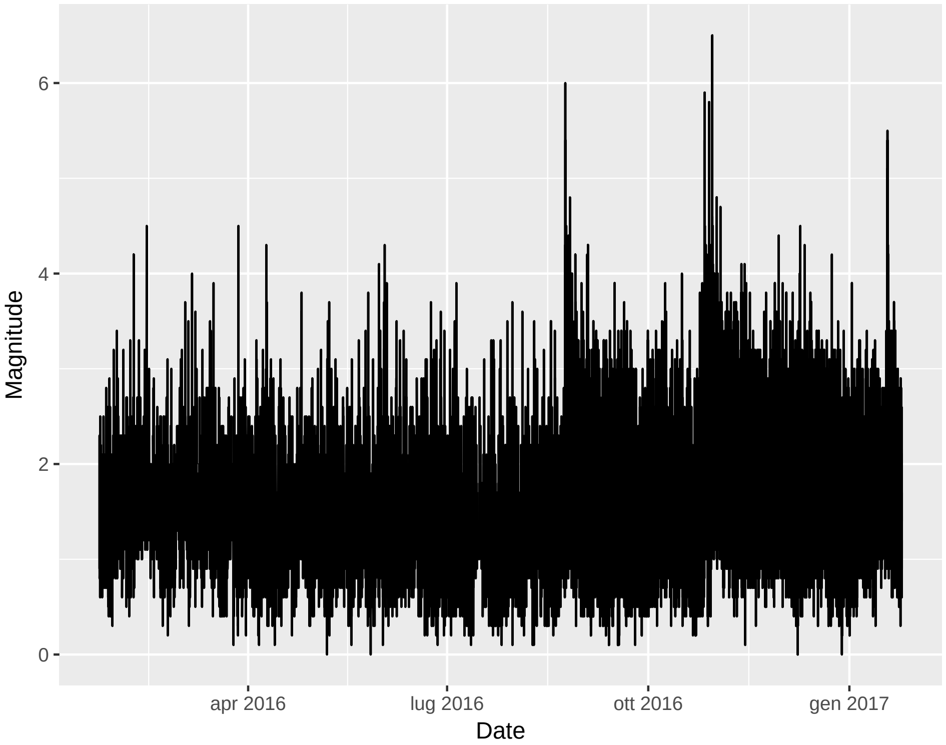

Figure 1.1: Time series of the Italian seismic events between 24^{th} January 2016 and 24^{th} January 2017, according to the INGV data. The number of registered shocks is 59190.

approach to the seismic data analysis is not new, many studies can be quoted as examples about the investigation of the magnitudes by using a rank-size approach (see [2, 54, 81, 104, 111, 119, 141]).

At the best of our knowledge, it is the first time that the Italian recent seismic events are analyzied in this way and that power laws are employed for proposing the quantification of damages form earthquakes. Specifically, the observed magnitudes along the considered period are transformed into costs by the means of the ZML and the UL calibrated with the mentioned catalogue. These cost indicators could be helpful in the definition of government policies with respect to the risk of natural disasters just like the seismic events.

A robustness check of the outcomes has to be performed in order to draw reliable conclusions. For this reason we run additional analyzes on two different datasets. One is an enlarged catalog which contains the Italian seismic time series from the 16^{th} of April 2005 to the 31^{st} of March 2017, downloaded from the INGV website. In April 2005 has been established a new network for seismic events registration, it is further improved by the time passing, but in order to avoid catalog incompleteness biases, we consider a $M_c = 2.5$ as magnitude lower bound, as in [115, 120]. The second differentiation of data used for another robustness check is elaborated in order to explore the space effect of the considered variable on the dataset that covers January 24^{th}, 2016 - January 24^{th}, 2017. We make new sub-catalogs by picking seismic events with epicenters in the eight adjacent provinces: Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo (with respective coasts). On each of them we perform the rank-size analysis. In this way we are consistent with the scholars that point the irrelevance of spacial effects for short time periods and small regions as [103]. Both the procedures addressed to validate the results of the main analysis give no signal of a remarkable weakness presence. Indeed, for the first case we have similar outcome obtained in the main analysis and for the second the difference between provinces is not significant, such that the main results can be considered valid.

In the next sections we give a detailed presentation of the data, the methodologies applied, the formulation of the costs, the outcomes of robustness assessments and presentation of the results with a wide discussion of them.

1.1 Catalog of the earthquakes

The magnitudes of the earthquakes occurred in Italy during the days between January 24^{th}, 2016 (first hours of the day) and January 24^{th}, 2017 (midnight), are part of the catalog here examined, see Figure 1.1. We include the earthquakes that have destroyed Amatrice on August 24^{th} 2016, with magnitudes equal to 6 and 5.3, that one with epicenter in Norcia with magnitude equal to 6.5 (October, 30^{th}, 2016), the one with magnitude equal to 5.9 of Ussita (October, 26^{th}, 2016) and the most recent occurred in Province of L'Aquila with magnitudes of 5.4, 5.1, 5.5 (January, 18^{th}, 2017).

The Italian National Institute of Geophysics and Vulcanology is one of the main Italian data provider for seismic events. From its web site (`http://cnt.rm.ingv.it`) we download the dataset and we take the magnitude definition that is characterized by different ways of transforming the registered seismic signals, range of magnitudes and scales, and the distance from the epicenters. In order to have more information about the definition of magnitudes, see [51].

The available data on the INGV web site is an enormous list of seismic events with information about depth and localization of the epicenters. The period that is under consideration here is one of the most active with 59190 observations, most of them are occurred in the center of Italy in about 6 months. For this study we examine the magnitudes and then the epicenters just in the final robustness checks.

We deal with the problem of catalog incompleteness because the shocks with highest magnitudes might hide many aftershocks, causing registration losses that are not easily quantifiable. In order to face this issue, we decide to truncate the catalog by restring it to the seismic events with magnitude not smaller than 3.1, following the steps of [28, 115, 120]. In so doing we obtain a dataset with 978 observations.

The main descriptive statistics of the data are reported in Table 1.1. In Figure 1.2 is reported the time series probability density function with the fit of a power law function calibrated on the earthquakes intensities. This kind of scaling behavior is investigated by many scholars (e.g. [58]) and for this reason it is expected.

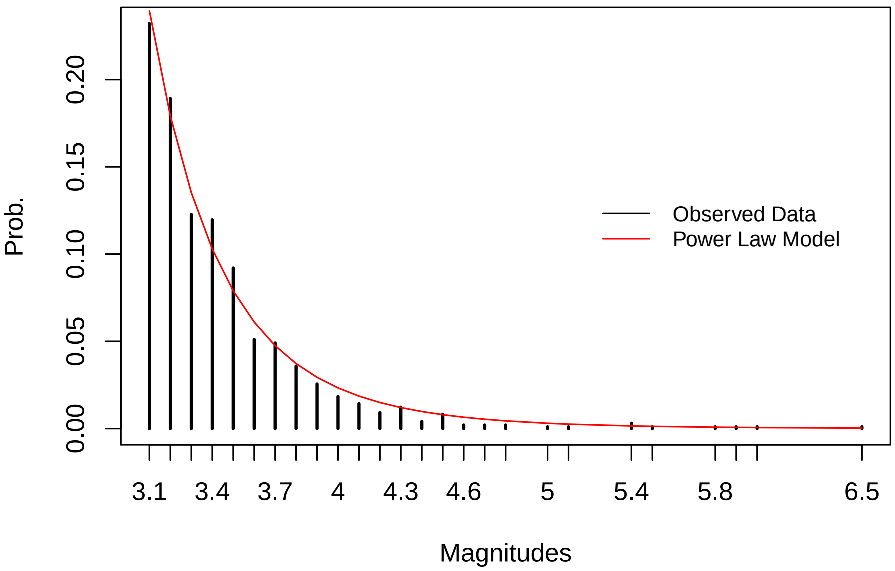

Figure 1.2: Probability of the earthquakes occurred between January 24^{th}, 2016 and January 24^{th}, 2017. The magnitudes are not smaller than 3.1. The model for the fit is a power law of the type $y = ax^b$. The estimations are $\hat{a} = 7428.58$ and $\hat{b} = -9.14$, with an R^2 of 0.99.

Statistical indicator	**Value**
Number of data	978
Maximum	6.50
Minimum	3.10
Mean (μ)	3.42
Median (m)	3.30
RMS	3.45
Standard Deviation (σ)	0.39
Variance	0.15
Standard Error	0.01
Skewness	2.67
Kurtosis	14.36
μ/σ	8.73
$3(\mu - m)/\sigma$	0.95

Table 1.1: Statistical features of the Italian shocks with magnitude not smaller than 3.1 occurred during the period: January 24^{th}, 2016 - January 24^{th}, 2017.

1.2 Methodologies and Cost Indicator Proposal

As announced in the previous section, the scope of Chapter 1 is to propose an aggregate cost indicator by transforming the magnitudes through rank-size laws. Indeed, we consider the magnitude as the size of the rank-size analysis.

An earthquake of a certain intensity can make a different level of damages according to the length of the seismic series precedes or follows and depending on the concentration of the epicenters on a given territory. Let us suppose that an earthquake of magnitude z is occurred at time t and that in the interval $[t - \Delta t, t]$, n foreshocks of magnitudes $z_1, ..., z_n$ are occurred. We transform z into $\tilde{z} = \eta(n, z_1, ..., z_n, \Delta t) \times z$ where $\eta(n, z_1, ..., z_n, \Delta t)$ is a parameter that increases with respect to $z_1, ..., z_n$ and n decreases with respect to the time window $[t - \Delta t, t]$, with Δt not smaller than 1. This conceptualization is helpful to conclude that if in a given zone, a certain number of foreshocks with remarkable intensity precede a main shock that occurs at time t, then the damages created by the last one are attributable to those caused by an isolated earthquake of magnitude $\tilde{z} > z$.

The parameter $\eta(n, z_1, ..., z_n, \Delta t)$ has to be determined in order to be employed in the proposal of an aggregate cost indicator. But in this case we assume $\eta(n, z_1, ..., z_n, \Delta t) = 1$, for each $n, z_1, ..., z_n, \Delta t$, that means no cumulative effects from foreshocks, namely each earthquake is treated as if it is unique and isolated. This setup gives back the lowest cost estimation and it is satisfactory for our study because it highlights the consequences of a weak anti-seismic policy anyway. This setup implies that $\tilde{z} = z$. From now on we use to z referring to $\tilde{z}$ without explicitly mentioning the distinction, just for simplifying the notation.

The catalogue is sorted in decreasing order by using the magnitude of earthquakes. In this way in the lower ranks there are shocks correspondent to the strongest events in term of magnitudes while in the higher ranks there are earthquakes with smaller magnitudes. In fact, at rank $r = 1$ is corresponding the highest registered magnitude while at $r = 978$ corresponds the shocks with magnitudes equal to 3.1.
We use the ZML and UL in order to explore the relation between the size-magnitude z and the respective rank. The two functions used are:

$$\tilde{z} \sim f_{ZML}(r) = \alpha(r + \beta)^{-\gamma}, \tag{1.1}$$

$$\tilde{z} \sim f_{UL}(r) = k\frac{(N + 1 - r + \psi)^{\xi}}{\big[N(r + \phi)\big]^{\lambda}}, \tag{1.2}$$

where α, β, γ are the parameters that have to be calibrated on the size data when eq. (1.1) is used, while k, ψ, ξ, ϕ, λ are those that has to be calibrated when we use the eq. (1.2). The parameter N corresponds to the number of observations that are 978 for this specific case.

The process for obtaining the damages economic indicator involves a transformation of the earthquakes magnitudes into respective costs. In sight of this, the restriction of the magnitudes lower bound at 3.1 makes sense because a seismic with lower intensity has a negligible probability of causing damages. Furthermore, we impose that the cost are positive and increasing for $\tilde{z} \geq 3.1$. The threshold $\tilde{z}$ depends on many factors, but the main one is the anti-seismic ability of the infrastructures that insist in a given area. When the residential and infrastructural system is particularly old or the normative on anti-seismic building is particularly defective , the $\tilde{z}$ can become very small, representing the considered territory as a very risky one.

Given that, we conceptualize two different cost indicators based on eqs. (1.1) and (1.2). We define $K_{\diamondsuit} : [0, +\infty) \to [0, +\infty)$ such that $K_{\diamondsuit}(z) = H(f_{\diamondsuit}(r))$, where $\diamondsuit = ZML, UL$. Quantity $K_{\diamondsuit}(z)$ is the cost associated to an earthquake with magnitude z when the best fit is performed through function $f_{\diamondsuit}$ and $H : [0, +\infty) \to [0, +\infty)$ increases in $[\bar{z}, +\infty)$ and it is null in $[0, \bar{z})$.

In order to find a critical magnitude $\bar{z}$ under the lights of rank-size analysis, one has to identify a respective critical rank $\bar{r}$ such that $z \leq \bar{z}$ if and only if $\bar{r} \leq r$. Given that we use eqs. (1.1) and (1.2) to perform the analysis, so we look for two critical ranks respectively identified as $\bar{r}_{ZML}$ and $\bar{r}_{UL}$.

Each earthquake included into the catalogue generates a certain economic cost whose aggregation is indicated with Γ. The maximum magnitude associated to a seismic shock that is ever registered is 9.5, it is the case of the Great Chilean earthquake in 1960. Here we name the maximum level of intensity as Z_{MAX} and we take it equal to 10 even if the empirical one registered in the reference period is 6.5 (see Table 1.1). Therefore, the cost indicators are:

$$\Gamma_{ZML} = \int_{\bar{z}}^{Z_{MAX}} C_{ZML}(z)dz = \int_{0}^{\bar{r}_{ZML}} H\left(\hat{\alpha}(r + \hat{\beta})^{-\hat{\gamma}}\right) dr, \qquad (1.3)$$

and

$$\Gamma_{UL} = \int_{\bar{z}}^{Z_{MAX}} C_{UL}(z)dz = \int_{0}^{\bar{r}_{ZML}} H\left(\hat{k}\frac{(N + 1 - r + \hat{\psi})^{\hat{\xi}}}{[N(r + \hat{\phi})]^{\hat{\lambda}}}\right) dr, \qquad (1.4)$$

which are respectively based on the fit of eqs. (1.1) and (1.2). The cost indicators are related to the value of $\bar{z}$ at ceteris paribus conditions. Of course, they increase as $\bar{z}$ decrease; they are null when $\bar{z}$ corresponds to the maximum level of magnitude, that represents the case in which the infrastructure is completely anti-seismic.

We design three different scenarios based on the idea that we have different realities to compare in order to understand which is the shape that is most conform to the economic costs dynamics:

(i)

$$H(z) = \begin{cases} \exp(z), & \forall z \in [\bar{z}, Z_{MAX}]; \\ 0, & \forall z \in [0, \bar{z}); \end{cases}$$

This is the exponential case. With this setup the highest magnitudes are strongly penalized in term of cost.

(ii)

$$H(z) = \begin{cases} z, & \forall z \in [\bar{z}, Z_{MAX}]; \\ 0, & \forall z \in [0, \bar{z}); \end{cases}$$

The linear case assigns, to each level of the intensity of an earthquake, homogeneous contributions to the cost.

(iii)

$$H(z) = \begin{cases} \ln(z), & \forall z \in [\bar{z}, Z_{MAX}]; \\ 0, & \forall z \in [0, \bar{z}); \end{cases}$$

In the logarithm case the function assigns lower values to the cost contributions that come from the lower magnitudes.

To easily recognize the prospected scenarios we add a subscript into the cost indicator Γ, such that, for example, $\Gamma^{(i)}_{ZML}$ is the Γ_{ZML} obtained when H is as in item (i).

1.3 Robustness Check

The results gained from the analysis exhibited in the previous sections need to be corroborated by two added investigations performed thanks to two dataset: a first more global than the reference one, and a second reduced in according with the territory analyzed. For the former case, we enlarge the dataset by expanding the time span taken in consideration. The wider catalogue allows us to consider an incompleteness level inferior to the 3.1 adopted for the main case, in fact for the global case we consider as fair a level of 2.5, in accordance with [115, 120]. We download from the INGV website data of 13239 seismic events recorded with a magnitude not smaller than 2.5, occurred in Italy between the 16^{th} of April, 2005 to the 31^{st} of March, 2017. The 16^{th} of April, 2005 is an important date for the INGV and it is taken as starting point because it is the moment in which the Italian earthquake survey has changed to be impressively improved. A summary of the statistical features of the catalog is presented in Table 1.2. Figure 1.3 shows the probability density function of the enlarged catalog with a power low fit that approximates very well the empirical data. It is comparable with Figure 1.2 made with the original dataset.

The exploration of the spacial effects is realized by running the best fit procedures on the sub-catalog of earthquakes with epicenters in provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo ([5] for epicenters' estimation precision). In these territories are happened the majority of 2016's seismic shocks, namely the 87% of the original dataset is comprised (see [128]). Therefore the catalog of the principal study is reduced to 849 observation in a time span that goes from January $24^{th}, 2016$ to January $24^{th}, 2017$. In [103] it is possible to find an example of study where is stated that taking into analysis small territories and short time period makes possible to neglect the spatial effects. So, our analysis of the sub-catalog is in line with this methodological approach and with the aim of building cost indicators of earthquakes' damages. Consequently it is not fundamental to take into consideration the spatio-temporal correlations among shakes. A fortiori ratione we can overlook it because the rank-size analysis is based on shocks intensity and the geological reasons of its generation have a secondary importance for the scope of this study.

In the spatial effects examination we decide for the same magnitude threshold of 3.1 that we adopt for the main analysis. It is useful to make the results comparable and to avoid the catalog incompleteness problem (see [75]). The main catalogue's descriptive statistics are presented in Table 1.4 while in Figure 1.6 there is the density function of the observed magnitude and the respective power law model. In Table 1.5 the best fit estimations results on eqs. (1.1) and (1.2) are presented. They are not so different from the outcomes that come from the main analysis. A graphical inspection is possible thanks to Figures 1.7 and 1.8 where are shown the best fit lines with the eqs. (1.1) and (1.2) respectively.

The fitting procedure presented in Section 1.2 gives outcomes summarized in the Table 1.3 and represented in Figures 1.4 and 1.5.

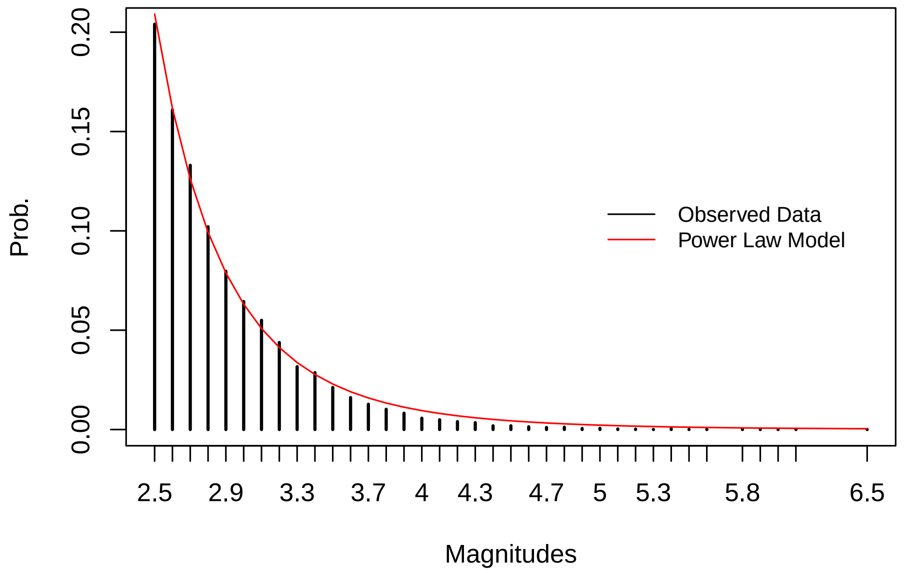

Figure 1.3: Probability of the earthquakes occurred between April 16th, 2005 and March 31st, 2017. The magnitudes are not smaller than 2.5. The model for the fit is a power law of the type $y = ax^b$. The estimations are $\hat{a} = 86.32$ and $\hat{b} = -6.57$, with an R^2 of 0.99.

1.3.1 Peculiar Catalog Features

We perform the best fit procedures described in Section 1.2 on the whole dataset, without any restriction to the magnitude that ranges from 0.1 to 6.5 in this case. It is realized because we want to explore some dataset features under a special data science perspective and without considering geophysical properties of the catalogue.

The statistical summary of the data here considered is presented in Table 1.6. The presence of outliers in the lowest and highest ranks strongly

Statistical Indicator	**Value**
Number of Data	13239
Maximum	6.50
Minimum	2.50
Mean (μ)	2.88
Median (m)	2.80
RMS	2.91
Standard Deviation (σ)	0.42
Variance	0.18
Standard Error	0.002
Skewness	1.89
Kurtosis	8.24
μ/σ	6.84
$3(\mu - m)/\sigma$	0.60

Table 1.2: Statistical features of the Italian shocks with magnitude not smaller than 2.5 occurred during the period: April 16^{th}, 2005 - 31^{st}, 2017

characterize the distribution of the data and mostly affect the performance of fitting the eq. (1.1) whose results are represented in Table 1.7 and Figure 1.9.

The chances of fitting the lowest and highest ranked values is incremented by using the eq. (1.2) (see [13]) as it is possible to notice from Figure 1.10. The parameters that have a fundamental role in this improvement are ψ and ϕ; they are shown in Table 1.7 for the fits with eqs. (1.1) and (1.2). When $\hat{\psi}$ is null the fit can capture the highest ranked magnitudes without flattering the final part of the curve as it is possible to see for the cases presented in Figures 1.5 and 1.8 and respective Tables 1.3 and 1.5.

The capacity of reaching the highest values is called the "queen" and "harem" effect in [12]. While, in [27] and [30] the authors have met an analogous situation for the lowest outliers. In these case, the extreme value with rank = 1 is called the "king" and the followings are called "viceroys".

The parameter ϕ in eq. (1.2) acts in the same way of ψ but to capture the effects of the lowest outliers. So, in presence of seismic events with very low magnitudes, the value of ϕ increases. Consistently with this idea the $\hat{\phi}$ is low for the case of Tables 1.3, 1.5, so when the catalogue is truncated and it is equal to 88.48 in the case here presented (see Table 1.7).

Eq. (1.1)	**Calibrated parameter**	**Value**
	$\hat{\alpha}$	9.48
	$\hat{\beta}$	68.80
	$\hat{\gamma}$	0.14
	R^2	0.98
Eq. (1.2)	**Calibrated parameter**	**Value**
	$\hat{k}$	0.88
	$\hat{\phi}$	9.52
	$\hat{\lambda}$	0.11
	$\hat{\psi}$	36951.95
	$\hat{\xi}$	0.30
	R^2	0.99

Table 1.3: Estimations realized with the fit of the eqs. (1.1) and (1.2) for the Italian earthquakes catalog that covers: April 16^{th}, 2005 - March 31^{st}, 2017 (N = 13239, magnitudes not smaller than 2.5). The value of the R^2 in both of cases is reported.

1.4 Results and Discussion

The time series of seismic events occurred between January 24^{th}, 2016 and January 24^{th}, 2017 by considering the magnitudes not smaller than 3.1 presents statistical features summarized in Table 1.1. Most of the shocks have intensity of 3.3 and the distribution is concentrated around the distribution center. Mean and median are different and it is expected due to power law behavior showed in Figure 1.2. The distribution is leptokurtic because of the outliers and it has a right-tailed shape.

The results in Table 1.8 are the outcomes of the best fit procedures run on eqs. (1.1) and (1.2) whose respective Figures 1.11 and 1.12 exhibit the excellent ability of the two models of representing the empirical data.
One of the peculiarities of the catalogue is the presence of shocks with very

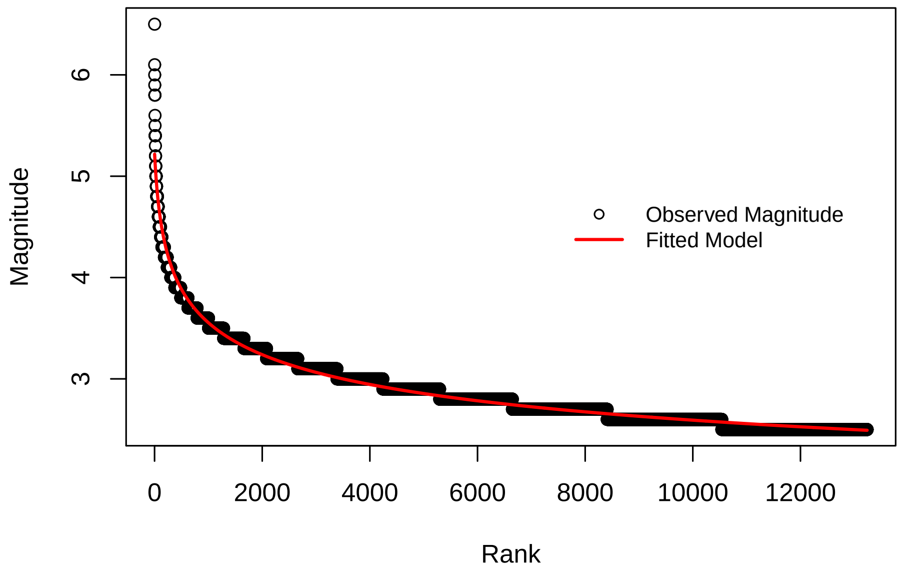

Figure 1.4: All the seismic events with magnitudes not smaller than 2.5 occurred in Italy during the years between: April 16th, 2005 and March 31st, 2017. They are sorted in decreasing order according to their magnitude and the ZML model is reported. See eq. (1.1).

high magnitudes (e.g. Norcia, the 30th of October 2016, magnitude 6.5 or Accumuli, the 24th of August 2016, magnitude 6) that are the outlier. Despite them, both the used model represented by eqs. (1.1) and (1.2) present the same great fitting ability on this dataset.
In order to strengthen the obtained results, in Section 1.3 we present the outcomes of the same best fit procedures, but running on a more global (longer time span: 2005-2017) and on a local catalog (provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo on a time span of one year: 2016 and early 2017). The local restriction of the dataset slightly affects the results because we exclude about the 15% of the observation from the original data and all of them have a magnitude ranked at low level apart one seismic event of magnitude 5. As a matter of fact, comparing the Tables 1.5 and 1.8 one can see the similarities, in particular for the ZML

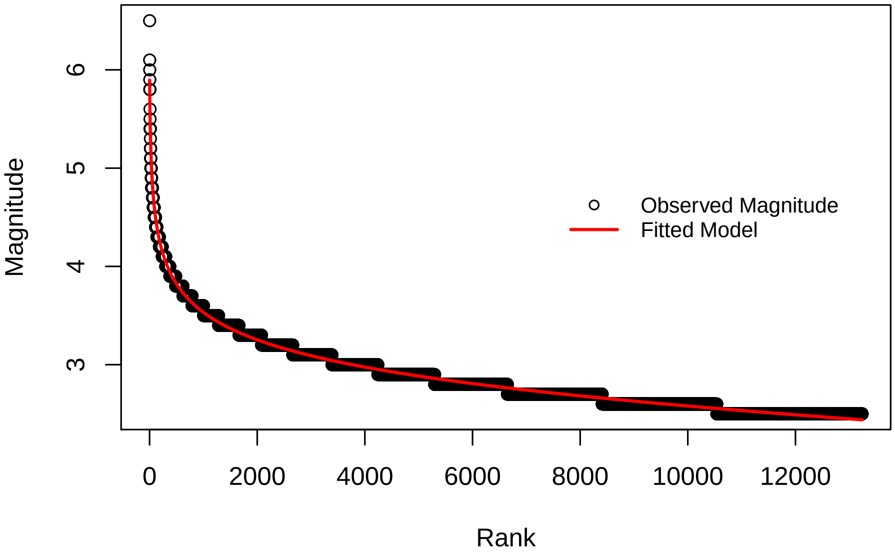

Figure 1.5: All the seismic events with magnitudes not smaller than 2.5 occurred in Italy during the years between: April 16th, 2005 and March 31st, 2017. They are sorted in decreasing order according to their magnitude and the UL model is reported. See eq. (1.2).

cases, while the UL's parameters estimations are lightly more sensitive to data variation. ZML's $\hat{\beta}$ is very close to zero and $\hat{\gamma}$ is very small for both the case of principal and local cases; it is proving the model tentative of capturing the effect of lower ranked magnitudes. Because of the exclusion of a shock with a magnitude of 5, we can appreciate the smaller value of $\hat{\alpha}$ in the case of ZML calibration in Table 1.5. These comments can be confirmed thanks to a visual inspection of the Figures 1.7 and 1.11 for the ZML cases and Figures 1.8 and 1.12 for the UL cases.

The comparison between the original catalogue that covers about one year and the case of the more global analysis whose time span is about 12 years highlights considerable differences. In the larger catalog there is a relevant increment of observations at higher ranks that is not proportional to the increased presence of magnitudes at lower ranks (see Figure 1.3). Tables 1.8

Statistical Indicator	**Value**
Number of Data	849
Maximum	6.50
Minimum	3.10
Mean (μ)	3.42
Median (m)	3.30
RMS	3.44
Standard Deviation (σ)	0.39
Variance	0.15
Standard Error	0.01
Skewness	2.75
Kurtosis	15.05
μ/σ	8.79
$3(\mu - m)/\sigma$	0.95

Table 1.4: Statistical features of the shocks with epicenters in the provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo, and with magnitude not smaller than 3.1. The reference period is: January 24^{th}, 2016 - January 24^{th}, 2017.

and 1.3 reports the parameters' calibrations of the two cases denoting a tiny loss in the fitting ability for the UL case in the reduced catalog.

As described in [13], equation (1.2), presented in the paper as universal law, is able to reach the highest values in the catalog thanks to ψ whose value depends on the weights of the lowest ranked elements. As a matter of the fact, the level of $\hat{\psi}$ in Table 1.3 is bigger than that one in Table 1.8 where the fit of eq. 1.2 is able to capture the effect of high ranked points without flattering the fitting curve at low ranks (see Figs. 1.12- 1.5). To catch the effect of the lowest magnitudes, the parameter mostly involved is ϕ, therefore $\hat{\phi}$ is expected to be large in case of the catalog with few restrictions as in Table 1.7 for the case of no restriction at all and Table 1.7 for a truncation at magnitude of 2.5.

In general, the goodness of fit returned from the applications of the Levenberg Marquardt nonlinear least square algorithm on the ZML and UL are very high, the R^2s line up between 0.98 and 0.99. The unique exception is given by the case of ZML application on one year catalog (January 2016 - January 2017) without any magnitude truncation, where the $R^2 = 0.93$. Anyway the dataset that produces such an estimation is not useful for the

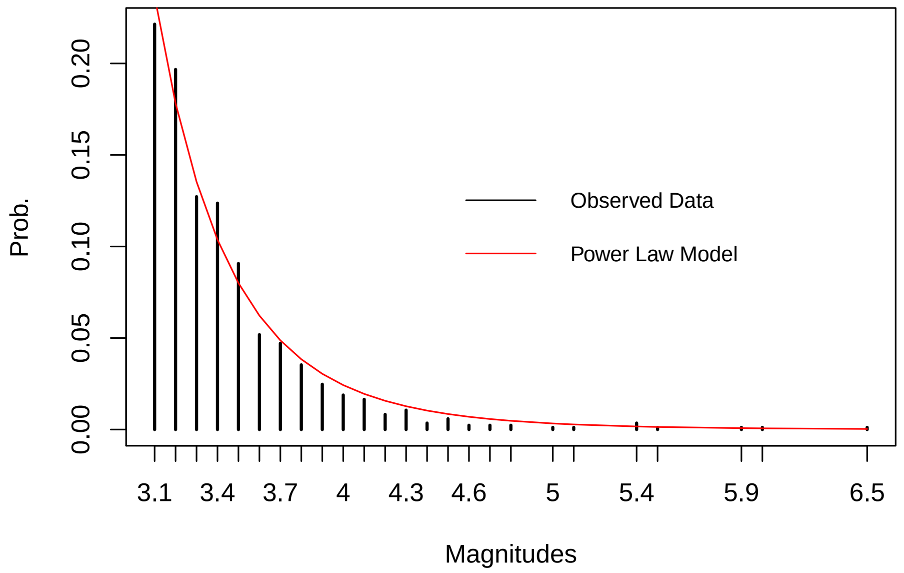

Figure 1.6: Probability of the earthquakes registered in the provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo between 24/01/2016 and 24/01/2017, with magnitudes not smaller than 3.1. The model for the fit is a power law of the type $y = ax^b$. The estimations are $\hat{a} = 5805.78$ and $\hat{b} = -8.93$, with an R^2 of 0.98.

cost evaluation due to catalog incompleteness problems. The impressive fit capacity of the ZML and UL can be immediately appreciated having a look to Figures 1.4, 1.5, 1.7, 1.8, 1.9, 1.10, 1.11 and 1.12.

Summarizing, eqs. (1.1) and (1.2) have a very close fitting ability for the original catalog, main object of this study, and for the dataset of the provinces involved in the seismic sequence of 2016-2017. So, the spacial effect results absent. The catalog that covers 12 years with a magnitude threshold at 2.5 is better represented by the UL than the ZML even if the R^2s result to be the same. It validates the calibration run over the original sample because the outcome are robust to time span used. The problematic of the catalog incompleteness is faced as suggested by the seismological literature by truncating the dataset at certain thresholds also suggested by scholars in

Eq. (1.1)	**Calibrated parameter**	**Value**
	$\hat{\alpha}$	6.07
	$\hat{\beta}$	0.00
	$\hat{\gamma}$	0.10
	R^2	0.98
Eq. (1.2)	**Calibrated parameter**	**Value**
	$\hat{k}$	9.50
	$\hat{\phi}$	0.00
	$\hat{\lambda}$	0.10
	$\hat{\psi}$	6749.18
	$\hat{\xi}$	0.02
	R^2	0.98

Table 1.5: Estimations realized by the fit of the eqs. (1.1) and (1.2) for the catalog that covers: January 24^{th}, 2016 - January 24^{th}, 2017 for seismic events with epicenters localized in the provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo (N = 849, magnitudes not smaller than 3.1). The value of the R^2 in both of cases is reported.

the filed.

The economic costs indicators are computed by integrating the eqs. (1.1) and (1.2) after transformation of different nature. Some integral can be calculated in closed form, so we have:

$$\Gamma^{(ii)}_{ZML} = \int_0^{\bar{r}_{ZML}} \hat{\alpha}(r+\hat{\beta})^{-\hat{\gamma}} dr = \frac{\hat{\alpha}}{1-\hat{\gamma}} \left[(\bar{r}_{ZML}+\hat{\beta})^{1-\hat{\gamma}} - \hat{\beta}^{1-\hat{\gamma}} \right] \tag{1.5}$$

$$\begin{aligned}\Gamma^{(iii)}_{ZML} = \int_0^{\bar{r}_{ZML}} \ln\left(\hat{\alpha}(r+\hat{\beta})^{-\hat{\gamma}}\right) dr = \ln(\hat{\alpha}) \cdot \bar{r}_{ZML} - \\ - \hat{\gamma} \cdot \left[(\bar{r}_{ZML}+\hat{\beta})\{\ln(\bar{r}_{ZML}+\hat{\beta}) - 1\} - \hat{\beta}\{\ln(\hat{\beta}) - 1\} \right];\end{aligned} \tag{1.6}$$

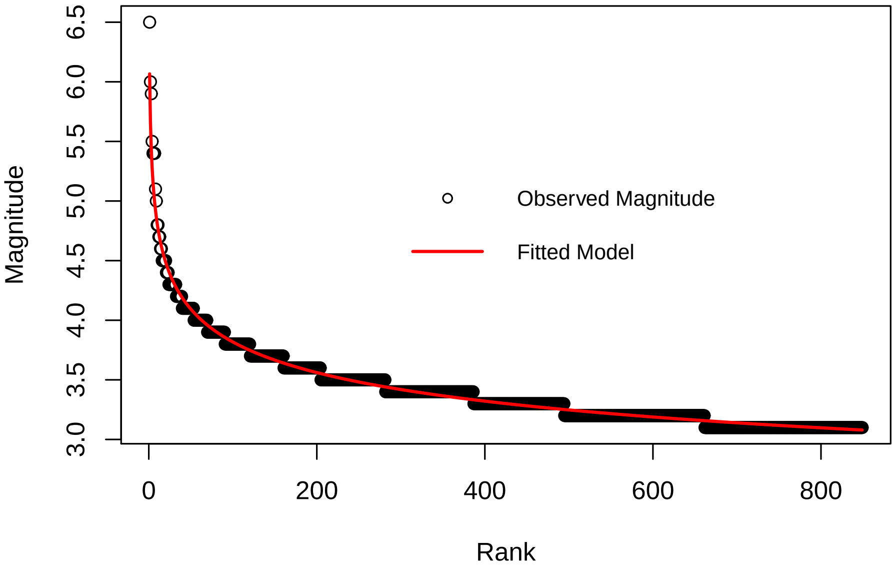

Figure 1.7: Seismic events with magnitude not smaller than 3.1 and epicenters in the provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo, occurred between January 24^{th}, 2016 and January 24^{th}, 2017. They are sorted in decreasing order according to their magnitude and the ZML model is reported. See eq. (1.1).

$$\begin{aligned}\Gamma_{UL}^{(iii)} = \int_0^{\bar{r}_{UL}} \ln\left(\hat{k} \cdot \frac{(N+1-r+\hat{\psi})^{\hat{\xi}}}{[N(r+\hat{\phi})]^{\hat{\lambda}}}\right) dr = \ln \hat{k} \cdot \bar{r}_{UL} + \\ + \hat{\xi}\left[-(N+1-\bar{r}_{UL}+\hat{\psi})\{\ln(N+1-\bar{r}_{UL}+\hat{\psi})-1\} + (N+1+\hat{\psi})\{\ln(N+1+\hat{\psi})-1\}\right] - \\ - \hat{\lambda} \cdot \left[\ln(N) \cdot \bar{r}_{UL} + (\bar{r}_{UL}+\hat{\phi})\{\ln(\bar{r}_{UL}+\hat{\phi})-1\} - \hat{\phi}\{\ln(\hat{\phi})-1\}\right].\end{aligned} \tag{1.7}$$

The remaining cases of Γs are evaluated thanks to standard numerical techniques. Therefore a Δr is defined in order to divide the segment $[0, \bar{r}]$ into S sub-segments. In this way one obtains $r_0 = 0, r_s = r_{s-1} + \Delta r, r_S = \bar{r}$

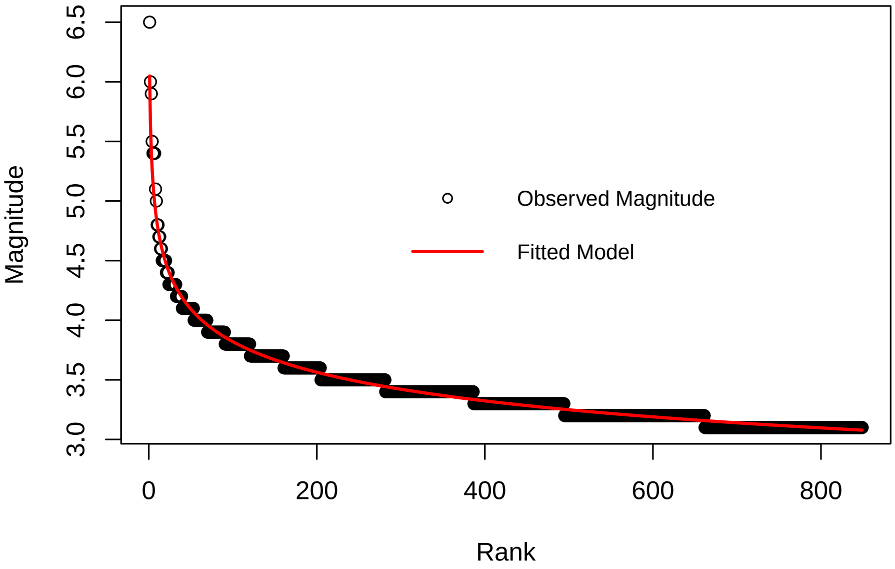

Figure 1.8: Seismic events with magnitude not smaller than 3.1 and epicenters in the provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo, occurred between January 24^{th}, 2016 and January 24^{th}, 2017. They are sorted in decreasing order according to their magnitude and the UL model is reported. See eq. (1.2).

so that:

$$\Gamma = \int_0^{\bar{r}} H(r)dr \sim \Delta r \cdot \sum_{s=1}^{S} H(r_s).$$

Each value of $\bar{r}$ coincide with a certain value of $\bar{z}$ that is defined as the threshold of magnitude that cannot causes damages because we assume that, at such a level, the housing system and infrastructure is able to withstand seismic shocks. In order to evaluate the cost indexes decay we can observe the graphs in Figure 1.13 where cases of Γ_{ZML}s and Γ_{UL}s are plotted as $\bar{z}$ varies. The lines are produced by adopting a $\Delta r = 0.01$ for approximating the eqs. (1.5), (1.6) and (1.7). It is immediately possible to notice that the decay behaviors is pairwise equal because of the almost identical ability of the

Statistical indicator	**Value**
Number of data	59191
Minimum	6.5
Maximum	0.1
Mean (μ)	1.57
Median (m)	1.50
RMS	1.66
Standard Deviation (σ)	0.54
Variance	0.29
Standard Error	0.002
Skewness	0.95
Kurtosis	5.19
μ/σ	2.92
$3(\mu - m)/\sigma$	0.41

Table 1.6: Statistical features of all the Italian shocks occurred during the period: January 24^{th}, 2016 - January 24^{th}, 2017.

Eq. (1.1)	**Calibrated parameter**	**Value**
	$\hat{\alpha}$	1450.52
	$\hat{\beta}$	12879.57
	$\hat{\gamma}$	0.65
	R^2	0.93
Eq. (1.2)	**Calibrated parameter**	**Value**
	$\hat{k}$	4.85
	$\hat{\phi}$	88.48
	$\hat{\gamma}$	0.16
	$\hat{\psi}$	0
	$\hat{\xi}$	0.22
	R^2	0.99

Table 1.7: Estimations realized with the fit of the eqs. (1.1) and (1.2) for the entire catalog here analyzed (N = 59191). The value of the R^2 in both of cases are reported.

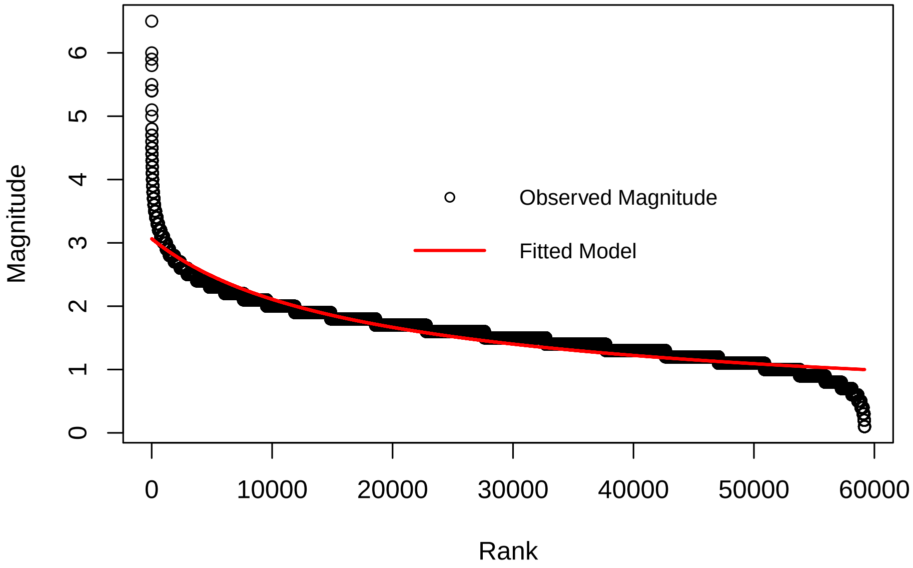

Figure 1.9: All the seismic events occurred in Italy during the years between January 24^{th}, 2016 to January 24^{th}, 2017. They are sorted in decreasing order according to their magnitude and the ZML model is reported. See eq. (1.1)

ZML and UL of replicating the real data. As expected the case in which in the damages are more relevant is the exponential one ($\Gamma^{(i)}$), on the other hand we get the opposite result for the logarithmic transformation of magnitudes into costs ($\Gamma^{(iii)}$). The decline of $\Gamma^{(ii)}$ and $\Gamma^{(iii)}$ is contemporaneous and converge to zero together, especially after $\bar{z} = 3.5$. $\Gamma^{(i)}_{ZML}$ and $\Gamma^{(i)}_{UL}$ have a fast converge zero before that $\bar{z}$ arrives around 3.7. After that value, the two lines decline more slowly than the other cases, evidencing a phenomenon that can be considered as lack of anti-seismic buildings.
It is remarkable to point also at the $\Gamma^{(i)}$s behaviors in the final part of the curves. Indeed, when $\bar{z}$ is about 5.7, there is a change in concavity that promptly collapses the economic damages to zero. This is considered an evidence in favor of policy that incentives the anti-seismic structures which

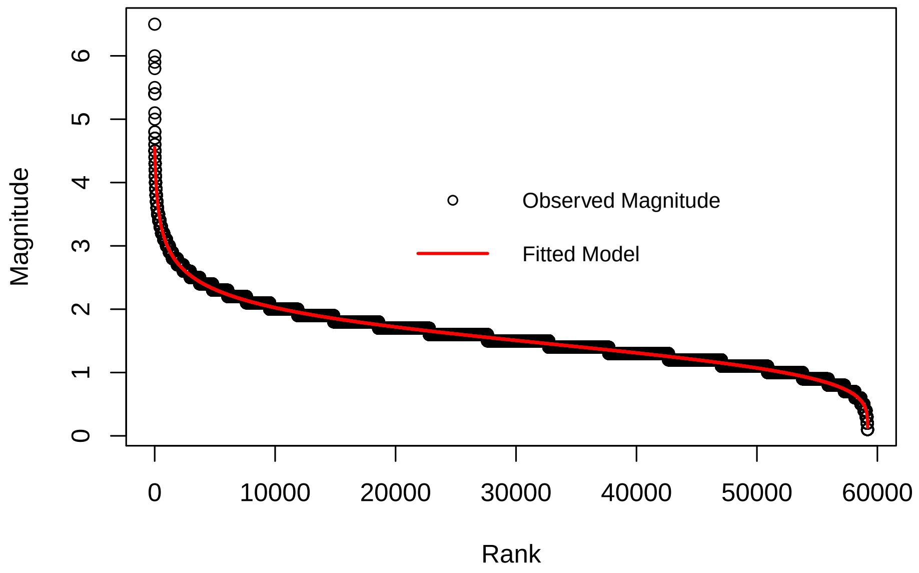

Figure 1.10: All the seismic events occurred in Italy during the years between January 24^{th}, 2016 to January 24^{th}, 2017. They are sorted in decreasing order according to their magnitude and the UL model is reported. See eq. (1.2)

are able to withstand to shocks above such a magnitude.

In Section 1.3 we present the robustness check of the fitting procedure of outcomes for ZML and UL. For the cost analysis we need to validate the results as well and the same economic analysis presented above is performed on the models calibrated in Section 1.3. Figures 1.14 and 1.15 present the costs decays for the ZML and UL calibrations on the three different catalogs used in this work: (a) for the original case, (b) local analysis and (c) is the global one.

The panels (a) and (b) of the costs estimated with eqs. (1.1) and (1.2) present a quite contemporary decrease with a very close aspect. The local cases, panels (b), are slightly different with respect to the first subplots, mainly because in the catalogue used is excluded a shock of magnitude 5.5, consequently the economic costs generated is reduced, leading to a feeble faster costs decays. The instance of 12 years dataset with truncation at magnitude 2.5, Figs. 1.14-1.15 panels (c), has a bigger number of imperceptible shocks that increase the damages if the territories has no basic anti-seismic ability. The main difference between the two subplots (c) is due to the fitting ability limitation of the ZML case, where the model is not able to interpret the outliers, reaching an estimated maximum magnitudes that is far from the observed one (compare Figs. 1.4 and 1.5). This fact leads the damages to be zero around a $\bar{z}$ of 5.4, while in Figure 1.15, subplot (c), the economic costs is zero near to 6.

Considering the first subplots of Figs. 1.14-1.15 that represent the economic cost dynamics of our interest, one can assert that the results are solid because they consistently reproduce the mechanisms of the studied object. Additionally, the second subplots do not affect the coherence of the results that maintain the same logic even if the study is performed over a bunch of provinces (for more information about spacial effects: [103]. Finally, the outcomes are not so sensible to the catalog completeness problem. Indeed, the evaluations here reported do not change that much when the dataset is truncated at 2.5 or at 3.1 apart for a marginal reduction of the economic damages considered for shocks with very low intensity.

Eq. (1.1)	**Calibrated parameter**	**Value**
	$\hat{\alpha}$	6.21
	$\hat{\beta}$	0.00
	$\hat{\gamma}$	0.10
	R^2	0.98
Eq. (1.2)	**Calibrated parameter**	**Value**
	$\hat{k}$	8.63
	$\hat{\phi}$	0.00
	$\hat{\lambda}$	0.10
	$\hat{\psi}$	6972.72
	$\hat{\xi}$	0.04
	R^2	0.98

Table 1.8: Estimations realized by the fit of the eqs. (1.1) and (1.2) for the Italian earthquakes catalog that covers the period: 24/01/2016 - 24/01/2017 (N = 978, magnitude not smaller than 3.1). The value of the R^2 in both of cases is reported.

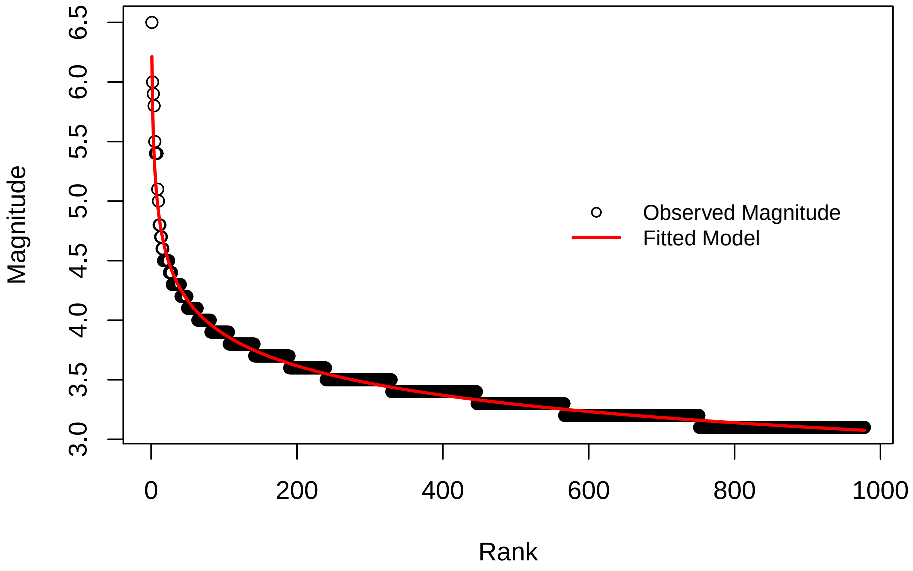

Figure 1.11: All the seismic events with magnitudes not smaller than 3.1 registered in Italy during the years between: January 24th, 2016 and January 24th, 2017. They are sorted in decreasing order according to their magnitude and the ZML model is reported. See eq. (1.1).

1.5 Concluding Remarks

In Chapter 1 we deal with three different datasets of the magnitude of the earthquakes. The main covers the Italian seismic shocks occurred between

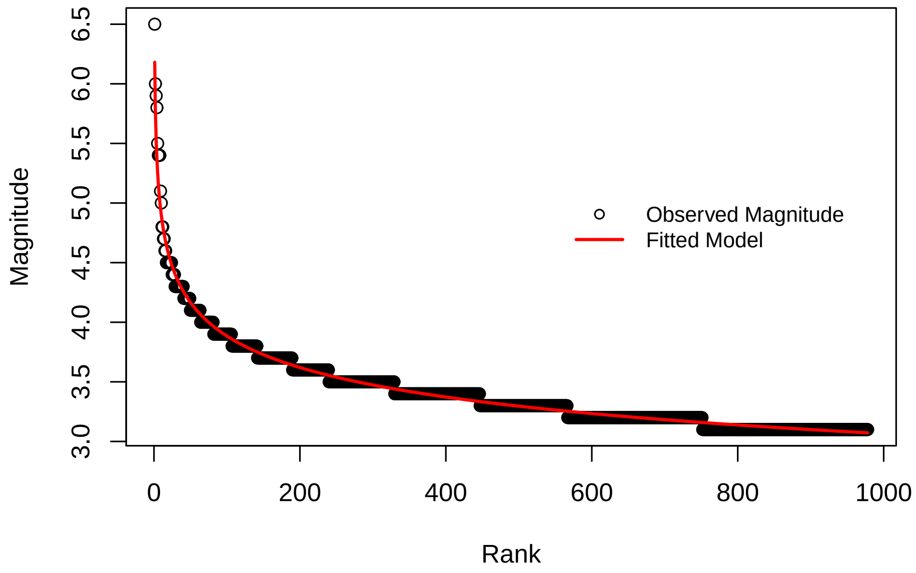

Figure 1.12: All the seismic events with magnitudes not smaller than 3.1 registered in Italy during the years between: January 24th, 2016 and January 24th, 2017. They are sorted in decreasing order according to their magnitude and the UL model is reported. See eq. (1.2).

January 24th, 2016 and January 24th, 2017 with $M_c = 3.1$. The second covers the same period and it has the same M_c, but the earthquakes selected have epicenters only in provinces of Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo. While the third contains the seismic events occurred between April 16th, 2005 and March 31st, 2017 and it has an $M_c = 2.5$. The last two catalog are mainly used to validate the findings of the analysis performed over the first dataset, specifically we treat them for exploring the problems of catalogue incompleteness and space effects.

Two different rank-size laws are presented: the Zipf-Mandelbrot law, equation (1.1) and the Universal law, equation 1.2. The parameters of the function are calibrated on the above described datesets thanks to the Levenberg Marquardt nonlinear least squares fitting with a brute force correction for avoiding the local minimums errors. Looking at the figures and the R^2s,

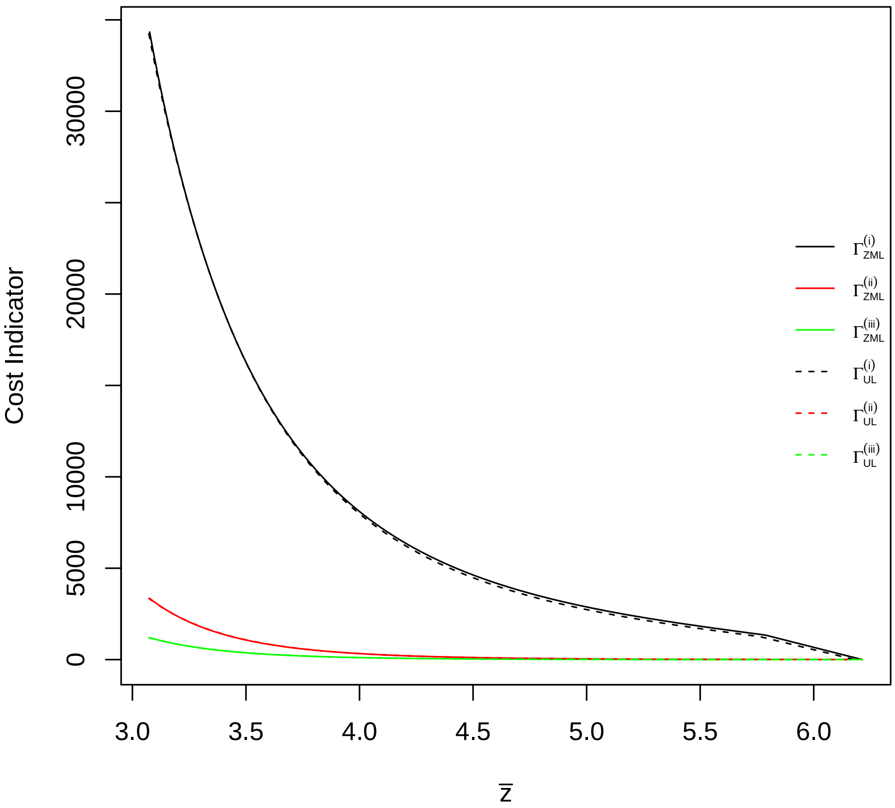

Figure 1.13: Comparison among eqs. (1.3) and (1.4) for the cases (i), (ii), (iii) as $\bar{z}$ varies. The considered catalog for the calibrations contains shocks with magnitudes not smaller than 3.1 and it covers the period: January 24^{th}, 2016 - January 24^{th}, 2017

it is possible to state that both the rank-size relationships are impressively capable of modeling the data with a small improvement of eq. (1.2) for the 12 years catalogue, especially when no-restrictions are considered on magnitudes (see Figs. 1.9 - 1.10).

In section 1.2 we propose three magnitude transformations for estimating the economic cost of shocks. In so doing we assume that there is a threshold in terms of the intensity of the seismic events that does not destroy the buildings. We state that such a magnitude limit can be increased by acting on building strategies by the means of policies. This type of conceptualization results coherent with the findings originated from the analysis on the original dataset (see Figs. 1.14-1.15, first subplots). Furthermore, the same analysis performed with the two other catalogs confirms the findings, authorizing us to tread as negligible the spacial effects and as solved the catalog incompleteness problem. As matter of fact, comparing the second subplots of Figs. 1.14-1.15 with the first, one can see that the mechanism represented has the same behaviors in term of shapes and decays. The same can be stated for the third subplots of figs. 1.14-1.15 that compared with the first subplots, do not show substantial differences apart from the fact that, for the case of larger dataset, we use an $M_c = 2.5$ that lightly increases the damages at left side of the figures, increasing the coherence with the logic of phenomenon.

Concluding, the evaluation of the economic costs of earthquakes evidently stresses that for reducing the damages caused by shocks it is necessary to increase $\bar{z}$, that represents the limit level below which the territories buildings withstand to seismic events (Figs. 1.13-1.14-1.15). To further distinguish dynamics of costs decay when $\bar{z}$ changes, we propose three different functional transformations of the magnitude.

All the findings of this chapter are addressed to suggest the adoption of a risk management plan for realizing the damages reduction here proved.

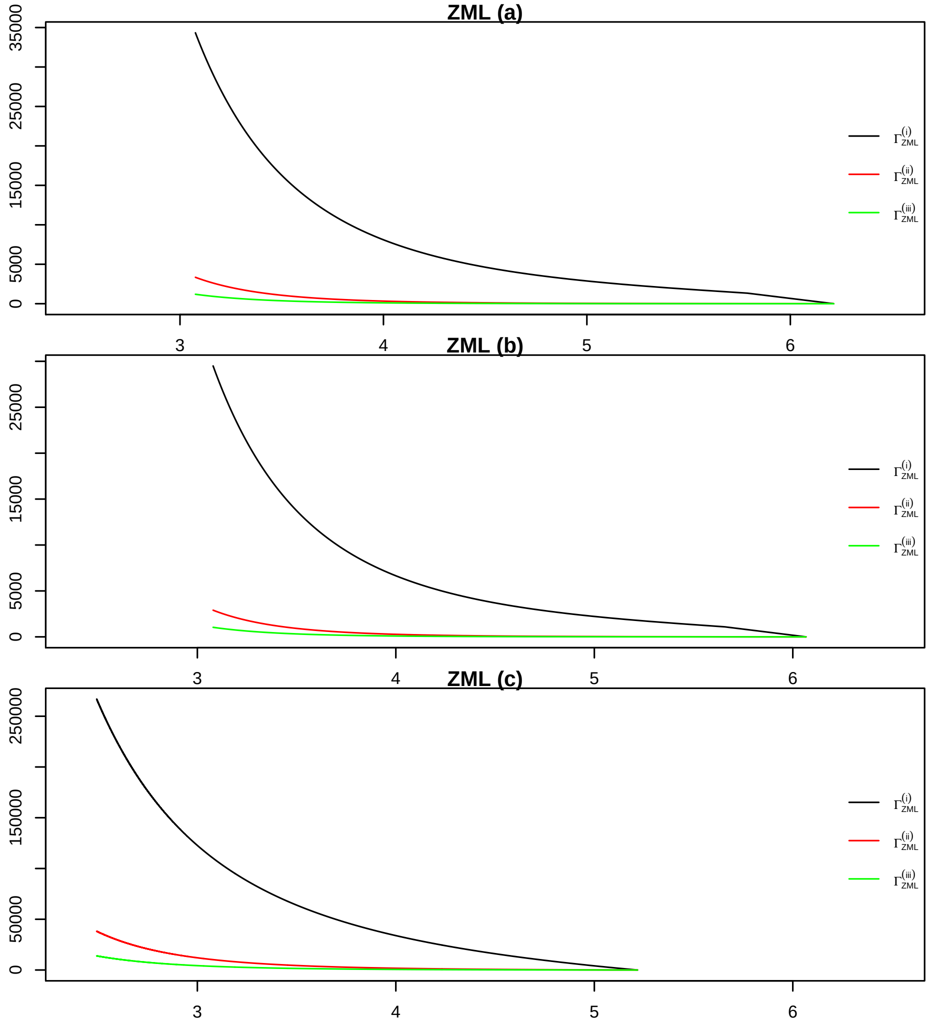

Figure 1.14: (a) Comparison of eq. (1.3) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 3.1 occurred between January 24^{th}, 2016 and January 24^{th}, 2017 in Italy.
(b) Comparison of eq. (1.3) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 3.1 occurred between January 24^{th}, 2016 and January 24^{th}, 2017 in Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo Provinces (comprised the respective coasts).
(c)Comparison of eq. (1.3) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 2.5 occurred between April 16^{th}, 2005 and March 31^{st}, 2017 in Italy.

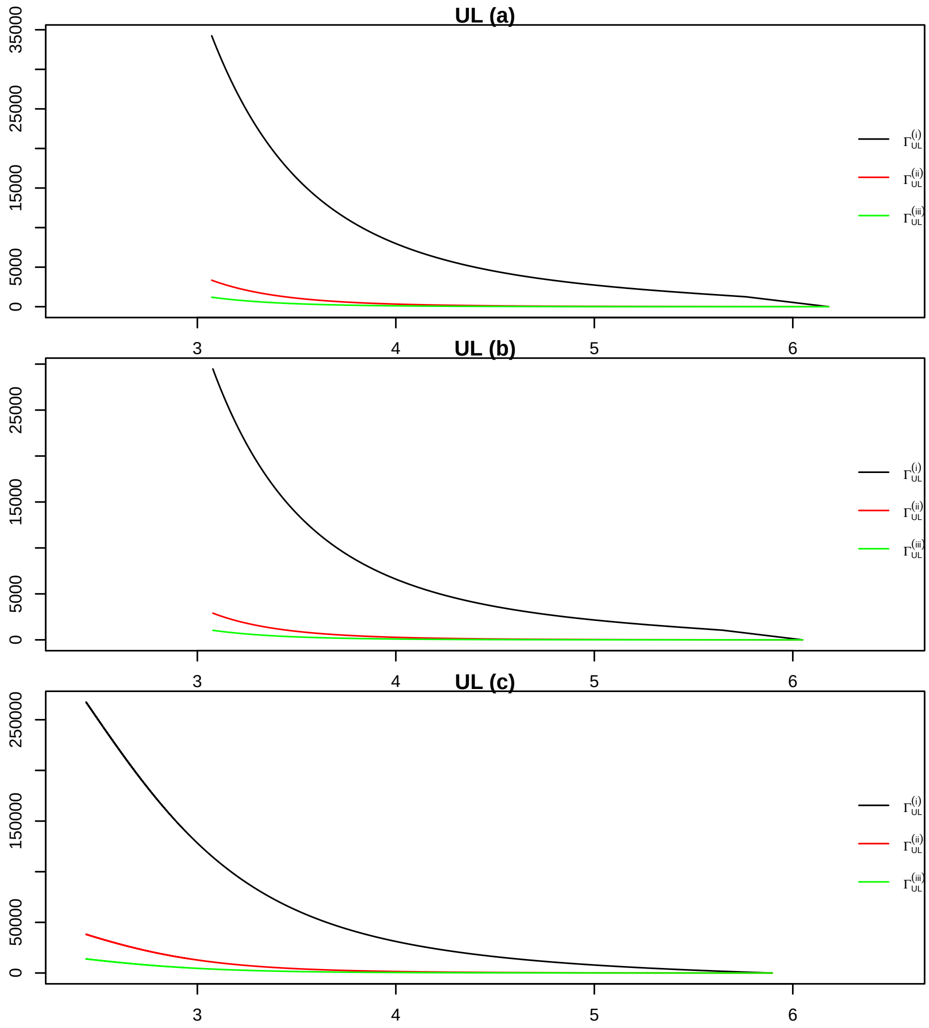

Figure 1.15: (a) Comparison of eq. (1.4) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 3.1 occurred between January 24^{th}, 2016 and January 24^{th}, 2017 in Italy.
(b) Comparison of eq. (1.4) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 3.1 occurred between January 24^{th}, 2016 and January 24^{th}, 2017 in Macerata, Perugia, Rieti, Ascoli Piceno, L'Aquila, Teramo, Terni and Fermo Provinces (comprised the respective coasts).
(c) Comparison of eq. (1.4) for the cases $(i), (ii), (iii)$ as $\bar{z}$ varies. The results are calibrated with a catalog that contains the shocks with magnitudes not smaller than 2.5 occurred between April 16^{th}, 2005 and March 31^{st}, 2017 in Italy.

Chapter 2

The socio-economic content and implications of political speeches – Part I: dataset building and global analysis

The main changes in schools of thought and political arrangements have been communicated to the public by means of speeches, whence by different rhetoric structures. The Presidents, trying to convince their own people, and others, about their opinions, have always used the words as the primary means (see [3]). The way to build the rhetoric structures is changed with time (about the changing in languages, see [135]). Evidences of these changes could be highlighted through text analysis in order to study classes of words as in [132].

The United States President is one of the most important people of the world and his speeches are often addressed to a wide audience. So, there is no doubt about the immanent relevance of the words pronounced by a President and it is expected that they bear a great influence on the overall economics and social contexts.

In this chapter, we deal with an analysis of the US presidential speeches in a larger sense than the mere focus on the meaning of the single words as it happens in exploring the topics present in a text (see [44]). Specifically our aim is to analyze the rhetoric structures of the speeches to gain insights on the way of creating speeches. For example, we denote a large set of words with low frequency or, differently, few words repeated several times, for each kind of talk.

In so doing, we are in line with language studies which investigate the structures of texts, speeches or languages (see e.g. [24, 10, 11, 4]). In our peculiar context, we follow the route traced by the studies on the connection between political speeches and Government policies (see [83] for the case of UK) and on the connections between US presidential talks and the surrounding environment (for the news' impact on financial market, see [36, 131]). It is worth mentioning [68] where the author identifies the main changes in the rhetoric of the President Inaugural Addresses and Annual Messages from George Washington to Bill Clinton. As we will see, the present study is radically different from [68]: (i) the present dataset is remarkably larger than the [68]'s one; (ii) the employed methodology is different: we use a rank-size approach, while the quoted paper adopts the General Inquirer for the specific assessment of words categories; (iii) for the target: we aim at giving a view of the structure of the speeches through the tokens frequencies while [68] pays special attention to the meaning of the words.

A key step of the research is the procedure for the creation of the dataset which contains about 1000 Presidents' speeches. Rough data is taken from the website: `http://millercenter.org` the 30^{th} of July, 2017, i.e. a set of 978 speeches, ranging from the *Inaugural address* of George Washington (1789) to the Donald Trump's speech *Address to Joint Session of Congress* (2017). As explained in the next sections, the number of speeches is reduced to 951 following data collection and treatment phases.

The study is performed through a rank-size analysis on the main part of the mentioned speeches. The size is defined as the (absolute or relative) frequency of words in each discourse, while the word rank has its position in the decreasing sorted list, so that rank 1 is that for the most often pronounced word in a speech. As introduced in the fist section, the rank-size analysis technique is a well-recognized method to explore the property of a large set of data when the data spans several decades and when error bars are not precisely defined due to sampling conditions.

We implement a best fit procedure thanks to the Levenberg-Marquardt Non-linear Least-Squares Algorithm in order to derive the ZML parameters for each speech. In so doing, we obtain a collection of best-fit parameters on the absolute frequencies as much as for the analysis over the relative frequencies.

To the best of our knowledge, this study is the first one dealing with the US Presidential speeches with a so large dataset. In this chapter, two classes of results are derived: the first one is associated to the dataset and the second class of findings relies on the rank-size analysis, whose parameters have peculiar meanings. The former class comes from the building procedure presented in a phase-wise including the pre-process phases, to ensure a comfortable replicability of it in other contexts and for further studies. In the respect of the latter case, we get non-linear regression on the ZML and the respective goodness of fit, hence arguing a common macro-structure among the speeches of the US Presidents.

The rest of the chapter is organized as follows. The next section provides a description of the dataset building procedure. In section 2.2 we propose a description of the collected data and successively we show the main features of the employed rank-size analysis. Finally, in sections 2.4 and 2.5 the results of the analysis along with a discussion of them are reported.

2.1 The dataset of the USA Presidents speeches: building procedure

This section is devoted to describe the dataset building process, namely we present the routine actions imposed by using R. The commands used to build the dataset are provided by the libraries *"xml2","rvest", "stringr", "xlsxjars", "xlsx"* along with their respective dependencies (see [140, 138, 139, 34, 33]).

The building procedure is divided into 13 phases.

In the first step, the considered website was visually examined in order to understand the structure of the contents. In particular, since one is looking for the presidential talks transcripts, it is important to find the pages where the addresses to the transcripts are listed. In the case of the Miller Center web site, the hyperlinks to each speech are dynamically showed in the following page: `https://millercenter.org/the-presidency/presidential-speeches`. Consequently, one has to inspect the HTML source code to find the objects of interest and to decide how to select them. In this case, it means that one needs to save the whole source HTML code to extract the links as showed in the next phase.

The second step is devoted to the first moves of the *web scraping technique* (see [56, 101] for the specific case of R). Such a technique is an automated computer science procedure for extracting information from websites through

a combination of dedicated commands. Thanks to them, it is possible to systematically access web pages in order to extract data of interest. The phases of a web scraping routine are twofolds: the first stage is characterized by the saving of the HTML source code of the web pages; the second stage consists in the extraction of the portion of the code where the needed information is reported. These actions are performed thanks to the functions: $read\_html(), html\_nodes(), html\_attr()$ and $html\_text()$. Such functions are available in the "rvest" library, which is employed to find systematically the links to the transcript of the speeches and other contents of interest.
So, in the second step, one grabs all the URLs (the acronym standing for: Uniform Resource Locators) of the speeches in order to prepare a list of links to be opened. The process of saving the addresses from the reference web page might lead to the occurrence of some errors. Such errors can be including mistakes produced by the web site creator. As an example, the links to the speeches pages could be reported by a different HTML identifier into the page, and this would lead to empty memorization. A control procedure is applied in order to face this problem; at the end of this phase, 978 addresses to the transcripts web pages were obtained.

The third step consists in the application of another web scraping routine on each page that contains the transcribed words. The list of links pre-loaded in the previous phase is treated to obtain transcripts of speeches, titles, dates, places of the statement, sources, and resumes of the speeches. This step is implemented through a *"for"* loop over the list of links that points to the pages where the speeches are stored. In each for's cycle, the web scraping routine is applied for the second time in order to read the HTML code of each web page pointed by the for running index. At the same time, one controls for possible discrepancies that could occur through an *"if"* statement inside the loop. This has to be done because sometimes the web pages where the speeches are presented could contain errors like: blank area where the corpus is supposed to be and/or the transcript is reported into another web page's section. Consequently, one has to control that the page has some characters into the space devoted to the transcript, then, in negative case, one has to look for the speech into another section. Doing so we realize that the web page containing the speech: "Campaign speech in Indianapolis, Indiana" stated by Herbert Hoover in October 28, 1932 [86], is one of the remarkable exceptions. Indeed, in the corresponding web page the discourse transcript is positioned in the section dedicated to the resume. So, to capture it, we use the same HTML selector used for memorizing the resumes of the talk, which is usually positioned in the top-right side of each web page.

At this stage, a visual inspection of the obtained results highlights that some rows of link's list point to the same pages of the speeches. Consequently, in the third step, some information would be downloaded twice. In order to manage such a case, a control on the title of the speeches is applied. In particular, we check for the duplication of titles and saved their position. In this way, it is possible to eliminate the respective positions into the variables used for each type of information (titles, resumes etc..). So there are 7 duplicated speeches: January 20, 2005: "Second Inaugural Address", April 27, 1961: "President and the Press", June 12, 1895: "Declaration of US Neutrality", December 6, 1892: "Fourth Annual Message", December 9, 1891: "Third Annual Message", December 1, 1890: "Second Annual Message" and December 3, 1889: "First Annual Message". After this further control, we have 971 stored transcripts.

The fourth phase is employed to manage the presence of typos in the inspected web pages of the Miller Center web site. The typologies of typing errors that are more relevant for the analysis are all those that contrast a correct division of the text into different tokens. Examples are the situations where the space between two words, number and word or punctuation and word is missing. Such typos can generate strings like: *"you.Therefore"*, *"10,000of"* or *"thePresident"* etc., and they impede the software to divide the text according to the adopted tokenization method.

The procedure for managing such typos is the following: the transcripts are firstly stored as a list of strings in a variable. One looks into each string to find all words divided by points without spaces like: " years.And ", " slowed.And ". These two elements are for example found in: "2016 State of the Union Address" [95]. The problem is solved by inserting a space between the points and the following word. With reference to the previous example the result of the corrections are: " years. And " , " slowed. And ".

Then, one solves the issue of numbers followed and preceded by words without space interruptions. An example of this typo can be found in the "First Annual Message" of December 6, 1981 [88]. There, the following exceptions occur: "June30", "in1881", "length3", "since1860", "of250". With the same method used in the previous phase, this typos are found and corrected as follows: "June 30", "in 1881", "length 3", "since 1860", "of 250".

Lastly, one manages the typos generated by two consecutive words merged without spaces, with the former entirely made by lower case characters and the latter made by the first character in upper case and the rest in lower

case. An example is in the "Inaugural Address" of March 4, 1925 [84], where the wrong token " ourConstitution" is transformed through the correction process in " our Constitution".
The procedure described in the fourth step is developed by employing regular expressions, which is a simplified method for searching patterns into strings by means of pseudo-coding languages (for a formal definition see e.g. [96]).

A further problem is related to the interactions of the President. Indeed, the President often talks in front of a wide and active audience. In such cases, it could happen that he is interrupted by applause, laughter, sung slogans or very loud screams at which the President could sometimes respond. These situations are reported into the speeches transcripts: the applause and laughter are sometimes reported between round or square brackets, while other kinds of contents like interactions between Presidents and audience are displayed after the specification of the speaker. A quite complete example of the described situations is given by the speech stated by Barak Obama: "Remarks in Eulogy for the Honorable Reverend Clementa Pickney" [94].

Thus, the fifth phase is devoted to remove the parentheses (square and rounded brackets) and their contents. For this aim, one needs to systematically access the individual speeches. Doing so, for example, there are cases in which the parentheses appear in the text but as typos (see e.g. the "Sixth Annual Message" of December 4, 1928 [85]. This type of exceptional error can be detected by registering the lengths of the parentheses content. For the analyzed dataset, the non-suspicious length of the parentheses content amounts to about 600 characters. In order to identify this limit as reasonable for a string length between two brackets, a visual inspection of all the parentheses content is performed. Above such a critical threshold, one can consider that the strings bounded by two parentheses do not constitute a content to be eliminated but, rather, a typo (this is the case of a missing closing bracket). A control is implemented by means of an *"if"* statement, to check if the eliminated pieces of text do not exceed the 600 characters threshold.

In this context, we need to mention also public events like press conferences, that are typically followed by questions from the public or by the journalists. The questions are denoted by an initial uppercase "Q" followed by a punctuation character or a space.
Another well established type of Presidential public meeting are the debates, which are characterized by a dialogue among candidates and/or journalists. The above-mentioned elements constitute noises for the analysis of the Pres-

idential statements; therefore, the sixth phase is dedicated to the amendment of the transcripts from the strings that do not come out directly from the Presidents, unique source of interest of this study. In this phase, some speeches with peculiar complexity (like the debates, where the rhetoric structure of the Presidential speech may be questionable, being also driven by the conversation flow) are directly removed. In particular, we remove 13 Debates and 1 Conversation from the original Miller Center database. To this end, we have looked for the words "Debate" and "Conversation" into the list of the titles and removed all the transcripts corresponding to the titles in which such strings appear.

By means of a *"for"* loop, the transcripts have been fathomed in order to find the presence of "Q" followed by punctuation or space, because such strings are clear signals of the presence of a question (so, out of the President rhetoric, as mentioned above). For consistency, all the text after a question has been removed; indeed, Presidents words are driven by the conversation and are not relevant in the analysis of the rhetoric structure of the speeches. An example is the Ronald Reagan's speech: "Speech on Foreign Policy" stated in December 16, 1988 [90]. Some exceptions arise. For example "Q." is often used into the transcripts to report names' abbreviations, like *"J. Q. Adams"* or *"Q. Tilson"*. Such cases are treated through a visual inspection of the removed texts.

Furthermore, in the Miller Center convention adopted to report the transcripts of the speeches, the most noticeable words pronounced by the public and reported are preceded by strings like: "THE AUDIENCE", "AUDIENCE", "AUDIENCE MEMBER" or "Audience", while the President statements are preceded by strings like: "THE PRESIDENT" or "President" sometimes followed by the President's surname (i.e. see the speech "Remarks at the Democratic National Convention" stated by Bill Clinton in August 29, 1996 [91]).

By using the regular expressions, it is possible to remove the reported public interventions, thus leaving just the words pronounced by the President. Such an amendment is done by eliminating the characters between the second (or third in case there were multiple speakers like auditors and journalists) speaker markers (for example "AUDIENCE") and the President markers (for example "THE PRESIDENT"). This means that one has to meet jointly (one after the other) the markers of both speakers in order to select the unnecessary text portions. Consequently, this process might lead to the removal of an excessive piece of text in the unlucky case of an error in the transcript.

For example, this is the case of the presence of a second speaker marker that is not followed by the President's marker once he takes the floor again, and the missing marker appears only later in the text. To manage such error cases, we proceed by further analyzing the texts candidate to be removed. In particular, we consider a threshold of 400 characters above which the selected texts are not cancelled. That is decided thanks to a visual inspection of all the portions of the selected texts. A remarkable example of the described exception is the speech by George W. Bush on September 3, 2004, entitled: "Remarks at the Republican National Convention" [92]. In that transcript, there are many audience interventions which are reported and marked with the string "AUDIENCE:" but after one of them there is a lack of marker that should have indicated where the President's words appear again. Thus one would eliminate a bigger portion of text that ends when the string "THE PRESIDENT:" is met again. In order to avoid the loss of large bunch of data, we decided to leave the selected pieces of text longer that 400 characters inside the analyzed sample. After all, the included words are so few that they cannot affect the final result; this was confirmed by visually inspecting the few exceptions which were found.
Sometimes, at the very beginning of the speech transcript, it is possible to find a string of the type: "THE PRESIDENT:", "The President:" or "The President" followed by his surname to mark the starting point of the President's words. This string is not captured by the described process because it is the very first and it is not coming after the intervention of other speakers. Therefore, at the end of this phase, one has to control for the presence of strings of that type at the opening words of the text, and has to eliminate them in affirmative case.

The seventh phase manages the situations in which the President delivered messages with his wife. In this case, we apply a control using a string of the type *"Mrs. "* followed by the President's surname. Then, to meet consistency, we remove all the words of the speech from the starting point of the intervention of President's wife.

At this stage, one has to check that the listed modifications do not reduce a given speech so much that it becomes not suitable for performing a consistent analysis (for exploring the implications of too short transcripts in applying ZML fit see [29]). The control of the suitable length of the speeches is the scope of the eighth phase. We eliminate the speeches whose resulting number of characters is less than 600. This threshold is identified by inspecting the number of characters' distribution for each speech.

After this procedure, we have eliminated five speeches: "Press Conference with Mikhail Gorbachev" (July 31, 1991), "Press Conference" (November 17, 1967), "Press Conference" (August 18, 1967), "Press Conference" (December 30, 1966), "Argument before the Supreme Court in the Case of United States v. Cinque" (February 24, 1841).

A further control for identifying speeches with the intervention of people different from the President has been next implemented. In particular, in some Press Conferences the questions are introduced by the name of the journalist followed by the name of the newspaper. In order to capture such cases, one controls for the presence of the string "Press Conference" in the titles and for the absence of the mentioned markers of the questions ("Q" followed by blank or punctuation). The unique observed example of this type is: "Press Conference in the East Room" (July 20, 1966, [87]). This speech is directly eliminated from the list. In fact, from a visual inspection, one can see that it is mainly made of words not provided by the President but rather by journalists.

The ninth phase is devoted to a last check of the outcomes of the previous steps. In particular, the string: "THE PRESIDENT" followed by punctuation or by the surname of the President (all upper or lower case) is searched into the speeches. The presence of such a string points to talks in which there is the intervention of the public or other speakers; such interventions did not appear in the previous phases because they are reported in the first line of the speeches. As an example, make reference to Obama's speech: "Address to the United Nations" delivered in September 23, 2010 [93]. In other cases, the string is reported at the beginning of the speech just for indicating the point in which the president is starting to speak. This control was run over the entire dataset and not just on the modified transcripts as for the previous phases.

The tenth stage is called *tokenization phase* (for a formal definition and some practical examples see [74]). Indeed, when the transcripts of the talks are stored into the cells, long strings of characters are memorized without any particular distinction. Yet, in order to work with the frequency of the words, the speeches have to be refined until they become a list of comparable units of analysis. Consequently, one needs to split these strings of characters (one for each speech) by making the procedure in R to be able to recognize single words in accordance with the used definition of token.
Specifically, with the R library "tokenizers" (see [100]), it is possible to invoke the command *tokenize_words*(). Such a command divides the text by

using the blank space as a separator and without taking into consideration the punctuation except for the decimal and thousand numbers separators. Furthermore it does not consider the apostrophes between words as a separator, like in the case of contracted form of verbs. As an outcome, we obtain that the variable containing the entire speech is transformed into a vector whose components are the words. Moreover, all the letters of the words are converted from upper case to lower case. In so doing, possible ambiguities due to the case sensitiveness of the words is forcefully removed.

In the eleventh step we implement a control for the speeches that are doubly reported into the same page section, hence leading to have doubled words frequencies. Sometimes it happens that those repeated transcripts do not have exactly the same words in common, but they differ for few terms (for a maximum of about 20, according to our empirical experience). This can be noticed by observing that some speeches have very few words with frequency equal to one. For this reason, the control for finding the double repeated transcripts is done by checking if the number of tokens that appear only once falls below a certain threshold. Thus, for each speech, we divided the number of words occurring just once, by the number of different words used, (see Figure 2.1); thanks to a visual inspection, one fixes the critical threshold at 20%.
The failure of this check pointed out to some technical inconsistencies of the website; one of the affected transcript is: ”Remarks Honoring the Vietnam Warś Unknown Soldier” of May 28, 1984 [89]. To solve this bug, we divided the frequencies of terms by 2, each time the control on the threshold was failing. But, in so doing, the exceptional terms appearing once in the double reported speeches reached absolute frequencies equal to 0.5. For this reason one has to add a further control for eliminating the residual tokens with 0.5 absolute frequencies.

The twelfth phase concerns the creation of a table type variable that has a number of rows equal to the number of different words used in each single speech and two columns: one for the tokens and another for the frequencies. Each couple is sorted out by decreasing order of frequencies. The tables of the speeches are labeled by the title and date of the speech.
To make the dataset exportable and ready to be processed, one collects and saves the data into a comma-separated values file (i.e. a *csv* file). An exportable matrix with 951 couples of columns and 3933 rows is then obtained. Each couple of columns is dedicated to the so sorted list of words and their respective frequencies for the individual speeches. In order to obtain a rect-

angular matrix, the number of rows of the matrix is uniformed to the maximum number of different words used across speeches, which is 3931. Of course many speeches have a lower number of different words, and the empty cells are filled by *NAs*, which point to a missing value indicator. After this, two rows are added to the top of the table: the first one is used to report the speaker name of the talk; the second one is adopted to show the titles with the dates embedded. Hence, the matrix has 3933 rows.
As already preannounced here above, the last phase concerns the export of that table into a *.csv* file. In so doing, the result is a dataset which is easily analyzable, – also with different programming languages; this goes in the direction of making this study reproducible.

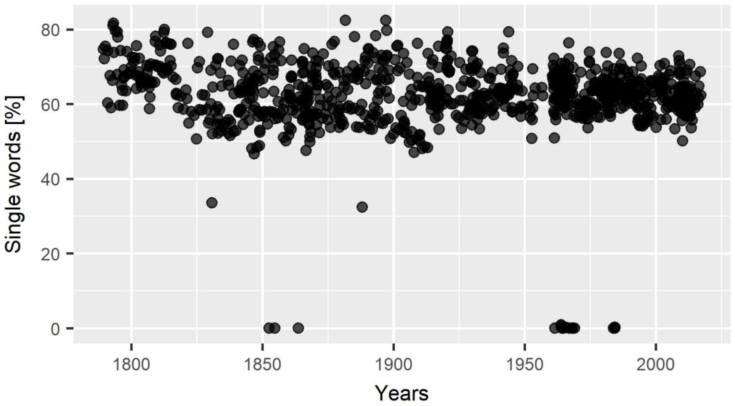

Figure 2.1: Percentage of words used just once on the number of different words used in each speech in a time varying representation.

2.2 The dataset of the USA Presidents speeches: description

This section contains the description of the dataset.

At the end of the process described in the previous section, one has a dataset of 951 Presidential speeches stated by the 45 Presidents of the United States, from George Washington to Donald Trump. The dataset covers a wide period of time: from April 30, 1789 to February 28, 2017. Due to Miller Center web site content, the number of speeches per President is different and it depends on criteria decided by the website owner. In the Table 2.1 there is the number of speeches per President reported in the dataset.

Miller Center website provides discourses stated in many occasions of the United State Political life: for example there are 57 State of the Union Addresses, 142 Annual Messages, 58 Inaugural Addresses, 20 discourses stated at universities or related to them, 18 speeches stated at National Conventions of Republicans or Democratic parties, 89 remarks pronounced by Presidents on salient topics and 567 other moments when the US Presidents have spoken to people.

All these declarations are collected as described in the previous section; furthermore, they are stored by organizing the words distribution for each speech. In this way, it is easy to apply the rank-size analysis in order to investigate the different rhetoric structures, as shown in the next section. Table 2.2 presents a statistical description of the speeches length in term of number of words per talk (second column) and in term of the different terms used in each speech (third column). The statistics will be commented upon in the result section. The discourse with the minimum number of different words used is the *"Message to Congress Requesting War Declarations with Germany and Italy"* in December 11, 1941, by Franklin D. Roosevelt; the one with the greatest variety of words is the *"Seventh Annual Message"* of December 3, 1907, by Theodore Roosevelt. It is interesting to note that the impressive amount of different words used in the latter message is related to the fact that, during this talk, the President Roosevelt has cited a part of another speech (Message to the Congress on December 5, 1905) and has mentioned events from the past, hence increasing the lexical richness of his speech.

2.3 Application of a Rank-Size law of Zipf-Mandelbrot type

The frequency of words in the speeches are the size of the rank-size analysis. Specifically, each talk transcript is stored into a table with terms and respective frequencies, so the rank $r = 1$ corresponds to the most repeated word of the speech. The tokens with lowest frequencies are stored in the positions corresponding to the highest ranks. The use of the plural is required because the terms with frequencies 1 and 2 represent generally the majority of the tokens in the discourses. Here, a best fit procedure to assess whenever the size-frequencies f and f_{rel} (absolute and relative respectively) might be view as a function of the ranks r is implemented. The considered fit function is the ZML:

$$f \sim g(r) = \alpha(r+\beta)^{-\gamma}, \tag{2.1}$$

where α, β,γ must be calibrated individually for each one of the 951 speeches. All fits are carried out through a Levenberg-Marquardt algorithm (see [65, 76, 69]) with no restrictions on the parameters. The starting points used in each estimation are provided through a non linear regression run over the same function but with a brute-force algorithm also known as grid-based searches, to avoid the dependence on starting parameters or getting stuck in local solutions. The same procedure is applied over the following formula, hence on the relative frequency of words.

$$f_{rel} \sim \frac{g(r)}{N} = \tilde{\alpha}(r+\beta)^{-\gamma}, \tag{2.2}$$

where N is the length of the considered speech and $\tilde{\alpha}$, β, γ are the parameters to calibrate.

The estimated parameters interpretations for the case of eq. (2.1) are the following: α gives information on the number N of words of the speech (see Figure 2.2), which is emphasized in Figure 2.3, and removed in the relative frequencies case of eq. (2.2). This aspect is discussed more in detail in the next section.

The parameter β, contains information on the higher ranked words. In particular the model reduces the weights of tokens with high frequencies when β increases. Moreover, if one does not have the presence of outliers, β is small.

For what concern γ, we expect that this parameter is close to 1, as it will be found; see results below. This parameter describes the concavity of the fitted models, whence it is informative about the distribution of words frequencies thereby giving an idea about their density. Indeed, if the magnitudes of the frequencies in the medium-high ranked words are high, then the calibrated $\hat{\gamma}$ is to be found small. Furthermore, γ is affected by the number of hapax legomena. If their presence is bulky with respect to the rest of the tokens, the model concavity will increase in order to capture the transition point between high ranks and low ranks.

This interpretation of the parameters is coherent with [20], and it will be also discussed later.

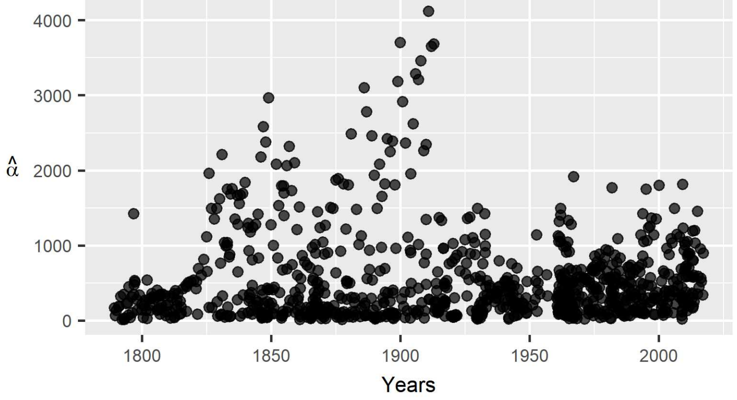

Figure 2.2: $\hat{\alpha}$ estimated on absolute frequencies with eq.(2.1) for each speech.

2.4 Results

The second column of the Table 2.2 contains the main statistical indicators of the length of the speeches in term of total used words in each speech. When looking at such statistics we can notice that the length varies considerably and its mean and median do not coincide. The positive skewness suggests a right-tailed shape, and the value of the kurtosis indicates a leptokurtic distribution. A similar situations is presented in the third column of Table 2.2 where the main statistics of the number of different words used in each speech are presented. The asymmetry is well identified by skewness value and confirmed by the different positional indicators. As for the previous case, we obtain a leptokurtic distribution.

The Figure 2.4 gives an idea of the talks lengths along the considered period of our sample and it is also informative about the years in which we have ”greater masses” of speeches by observing the density of the points. The recent years denote a change of behavior, indeed after the lack of speeches presented around the 1960 there is a noticeable concentration of the points. In a number of cases the lengths of the speeches and the variety of words are considerably high, probably due to the presence of very long speeches in which the President is reading some other documents or is quoting other talks. In Figure 2.1, it is possible to note the pattern of the percentage of words used only once per talks. We can note also a slightly decreasing trend that is confirming the effort of reducing the different words used or

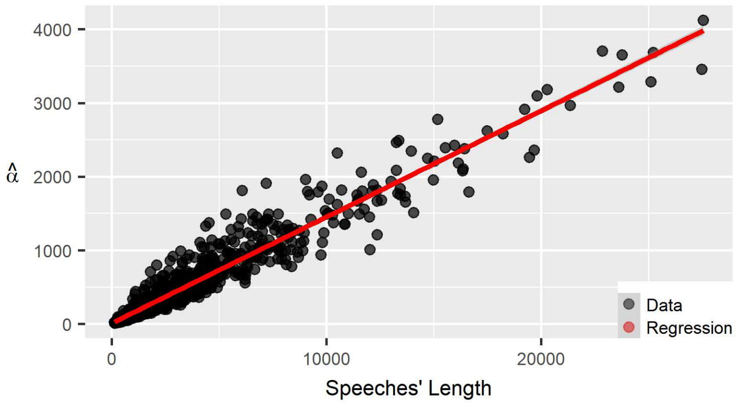

Figure 2.3: Relationship between $\hat{\alpha}$ and length of speeches in term of total number of words used per speech

a contraction of the dictionary richness. Anyway the speeches are mostly characterized by the presence of words pronounced only once. Such words populate the tails of each Zipf distribution and provide a characterization of them.

The words at ranks [1-10] generally belong to a group of largely pronounced tokens and in Table 2.3 we present the most frequent words fallen into that ranks range. In particular we report the probability of having such terms into the firsts 10 ranks of each speech.

The best fit procedure on eq. (2.1) and eq. (2.2) is performed speech by speech and a visual presentation of the goodness of fit measure for the second equation is reported in Figure 2.5. The main stats of R^2s for both the equation are presented in Table 2.5.

From the analysis of the absolute frequencies, some facts emerge.

A visual inspection of $\hat{\alpha}$ in Figure 2.2 shows a remarkable trend in the first years. Figure 2.3 evidences the positive correlation between $\hat{\alpha}$ and speeches lengths N, while Figure 2.6 shows that the correlation between $\hat{\alpha}$ and number of different tokens used in each talk is still present, but is less linear than the previous one. The dependence of $\hat{\alpha}$ on N represents a bias for the analysis of the results and a supportive argument for studying also relative frequencies. In fact, such a dependence disappears in calibrating the parameters with eq. (2.2), i.e. by taking into consideration the relative frequencies of the words, as shown in Figure 2.7. The calibrated parameters on relative frequencies are reported in Figures 2.8,2.9,2.10. A statistical summary of estimated parameter values is reported in Table 2.4; the goodness of fit measures is reported in Table 2.5. The salient cases of $\hat{\tilde{\alpha}}$ are presented in Figures 2.11, 2.12. The last one is associated to Ronald Rengan's speech titled "Remarks on the Air Traffic Controllers Strike", which is very short, thus having such a small $\hat{\tilde{\alpha}}$. We need to say that the original transcript was longer; we remove them as discussed in the second section. Thus, the parameter $\tilde{\alpha}$ can be viewed still as an indicator of the highest relative frequency in each speech, even if its magnitude is mitigated by the relationships with the other parameters.

The calibrated $\hat{\beta}$ gives an indication on the differences among the frequencies within the various speeches. Given that, the biggest differences are between the low ranked words as well known by [147, 146]. The $\hat{\beta}$s are showing the behaviour of the frequencies at the lowest ranks, thus on the most common words (see Table 2.3). An evidence of the feature of $\hat{\beta}$ is provided from establishing the differences between words' frequencies at consecutive ranks within each speech. Indeed, by summing for each speech the first 5 differences originated by the 6 most repeated words and comparing them to the $\hat{\beta}$s, one has Figure 2.13. Such a figure shows the decay of $\hat{\beta}$ with respect to the differences in frequencies within each speech. The graph is confirming that high level of $\hat{\beta}$ is corresponding to tiny differences between top six repeated words within speeches (see also Figure 2.14) while the converse occurs for the low level of the parameters (see Figure 2.15). This occurrence can be interpreted also in terms of the rhetoric structure, by asserting that when $\hat{\beta}$ is large, the words of the speech have a more homogeneous distribution along the highest ranks, hence pointing to the presence of a "rich club" of outliers at the low ranks.

Another interesting feature that emerges from the best fit procedure on eq. (2.2) is the correlation between $\hat{\tilde{\alpha}}$ and $\hat{\beta}$ (see Figure 2.16). The joint evaluation of such calibrated parameters gives then information on the magnitude of frequencies at low ranks.

The concavity of the fitted curve related to eq. (2.2) is mainly affected by $\hat{\gamma}$. Therefore this parameter is informative about the decay of the model passing by the low ranks to the high ranks. The $\hat{\gamma}$ is peculiar of each speech, changing the focus of the hyperbolas in agreement with the features of the talks. Consequently, it is the parameter that mostly affects the areas under the fitted models, which is reduced when $\hat{\gamma}$ increases (see Figure 2.17 for a graphical representation of the relationship between areas and $\hat{\gamma}$). The areas are calculated by computing the following integral over each model characterized by $\hat{\hat{\alpha}}$, $\hat{\beta}$ and $\hat{\gamma}$, where the i is the indicator of the i^{th} speech, so that $i = 1, \ldots, 951$:

$$A_i = \int_1^{r_{max,i}} \hat{\hat{\alpha}}_i (r+\hat{\beta}_i)^{-\hat{\gamma}_i} dr = \frac{\hat{\hat{\alpha}}_i}{1-\hat{\gamma}_i} \left[(r_{max,i} + \hat{\beta}_i)^{1-\hat{\gamma}_i} - (1+\hat{\beta}_i)^{1-\hat{\gamma}_i} \right] \quad (2.3)$$

A_i is the i^{th} area corresponding to the model calibrated over the i^{th} talks, while $r_{max,i}$ is the highest rank referred to the i^{th} transcript. Refer to Figure 2.18, where it is possible to notice the propensity of eq. (2.2) to converge to 1, giving an idea of the ZML capacity of being a density function in this specific case of the analysis of relative frequencies.

Notice also that the $\hat{\gamma}$ is low (high) when in the speech there is a more or less evident transition from low ranked words to high ranked words; see the two cases in Figures 2.19, 2.20.

Under an evolutive perspective, the most pronounced words for each speech have a constant decreasing rate along the years (see Figure 2.23). This result means that the repetition of a single word tends to reduce with time. This occurrence goes hand in hand with a global reduction in the differences between the frequencies of the words within the speeches, as shown in Figure 2.21. Considering that the most pronounced tokens are mainly conjunctions, articles and preposition (see Table 2.3), this phenomenon could be interpreted as the growing need in time of making simpler syntax of the sentences. Another useful hint of this fact is that the number of single words used (see Figure 2.1) are slightly decreasing along the years. So, the global tendency is to use less single words and propose simpler sentence structures.

2.5 Discussion and Conclusive Remarks

The samples of the presidential speeches analyzed in this chapter is one of the most complete in the literature. It is constructed under consistency criteria in a phase-wise form (see Section 2.1); finally it contains 951 talks over a span of about 228 years for all the US Presidents. The source of the data is the Miller Center website; data have been retrieved at the end of June 2017.

The use of rank-size laws with sizes given by the frequencies through the Zip-Mandelbrot law (see eq. (2.1) and eq. (2.2)) gives the opportunity of analyzing the rhetoric structure of the transcripts. More specifically, the method

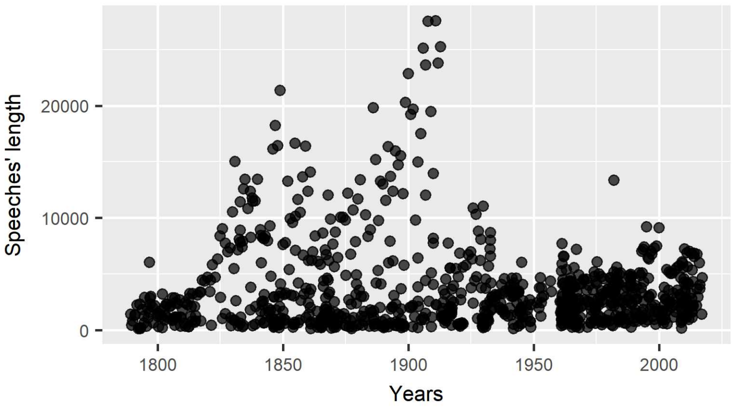

Figure 2.4: Length distribution of the speeches in term of total number of words used per speech over the years

allows us to observe changes into the rhetoric framework without being drastically affected by the changes in usages and significance of terminologies that occurred during the years. Indeed such an objective analysis of the frequencies disregards the meanings of the words and focuses on the exploration of the speeches frameworks. Undoubtedly, over the 228 years hereby considered, the American language has changed a lot as well as the "society" to which the speeches are aimed; thus, the convergence toward a common scheme in producing talks could be even explored searching for "universal behavior". Thus, as hoped, Eqs. (2.1) - (2.2) show a spectacular capacity of fitting the transcript, with values of R^2 always around 1 (see Table 2.5 and Figure 2.5). Further evidence of the impressive capacity of such a fitting can be noticed with a visual inspection of Figures 2.11,2.12,2.14,2.15,2.19,2.20. Therefore the selected approach does not reveal compromising weaknesses, it can be considered rigorous, and recommended in further works.

The calibrated parameters on eq. (2.2) are presented in Figures 2.8,2.9,2.10. It is possible to observe some changes along the years by visual inspecting $\hat{\tilde{\alpha}}$, $\hat{\beta}$ and $\hat{\gamma}$. Such calibrated parameters are used to resume some features of the speeches structures.

The $\hat{\tilde{\alpha}}$ of eq. (2.2) has a slightly increment in volatility during the last years with a high concentration of outliers after the 1960s. Considering the fact that $\tilde{\alpha}$ is giving an indication on the relative frequencies of the most often used words, the meaning of the related behavior along the years can

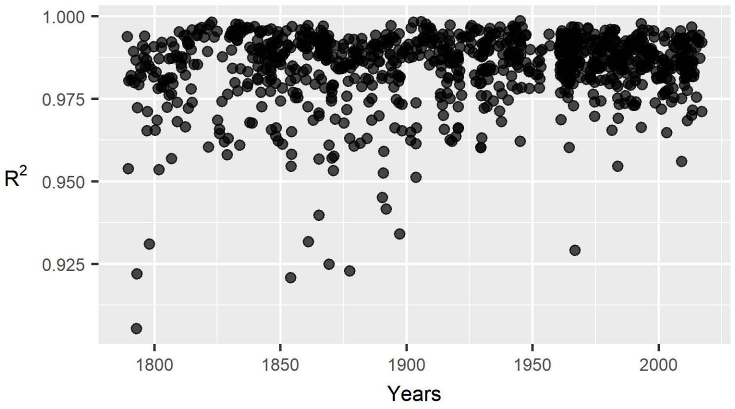

Figure 2.5: R^2 for each fitted speech on eq.(2.1) over the years.

be interpreted as an upcoming of irregularities in the use of words.

The analysis of $\hat{\alpha}$, so the parameter estimated with eq. (2.1), whose behavior is reported in Figure 2.2, leads to the assessment of two remarkable trends as regards the length of speeches during the years between 1800-1850 and 1850-1900. As we have said before, this outcome is grounded on the fact that the parameter α can be considered as a proxy for exploring the number of words employed in the speech.

The $\hat{\beta}$ has a similar behavior to that of $\hat{\hat{\alpha}}$, as can be seen from Figure 2.16. Indeed, its points are quite homogeneously distributed between 0 and 1 until 1900 when $\hat{\beta}$ starts to rise with a contemporaneous increment of the volatility (see Figure 2.9).

The β's increment when the differences between frequencies at low ranks are decreasing (see Figure 2.13) helps us to conclude that after 1900 the words frequencies distributions are converging toward more homogeneous distributions. A further confirmation of this is given by the areas delimited by the models and computed through eq. (2.3). As it is possible to deduce from Figure 2.22, there is a feeble positive trend, combined with a reduction in variability. Furthermore, there is a clear decreasing trend in the relative frequencies of the most often used words of each speech (see Figure 2.23), which reinforces the results and interpretations about $\hat{\beta}$.

From Figure 2.10, the calibrated $\hat{\gamma}$ appears to be quite stable in terms of trend and distance from 1. In most cases, such a parameter assumes a value around 1. When $\hat{\gamma} \geq 1$, then one can assert that there is a steeper decay

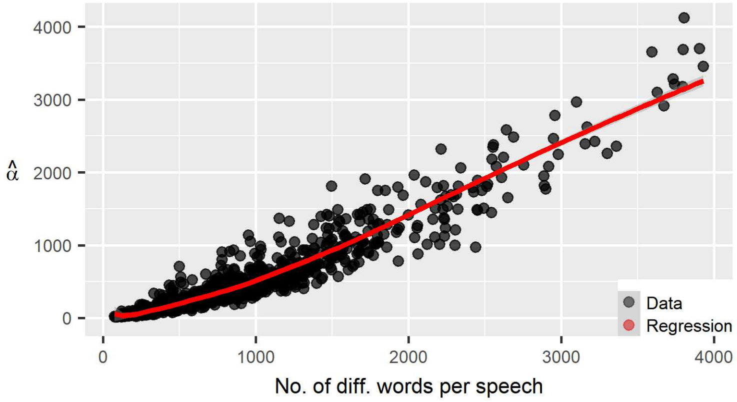

Figure 2.6: Relationship between $\hat{\alpha}$ and number of different words used in each speech (each word considered just once per speech)

of the data in the rank-size plot. Figure 2.24 assists in visualizing that the distribution of the $\hat{\gamma}$'s is asymmetric and the left tail is a bit longer than the right tail, giving a further indication of the tendency toward President producing "more homogeneous talks", in our sense.

Finally we can assert that the speeches structures exhibit in general a sort of common framework, with a specific proportion of words. Consequently, this means that the typology of the rhetoric involved in the political public speaking is identifiable. The method here applied is informative and the robust capacity of fitting provides a certain confidence in reaching a conclusion.

All these elements are opening the door for further studies. One of the most prominent proposal for future research concerns the assessment of the stochastic properties of the rhetoric of the speeches to perform some forecast in the structure of future presidential speeches.

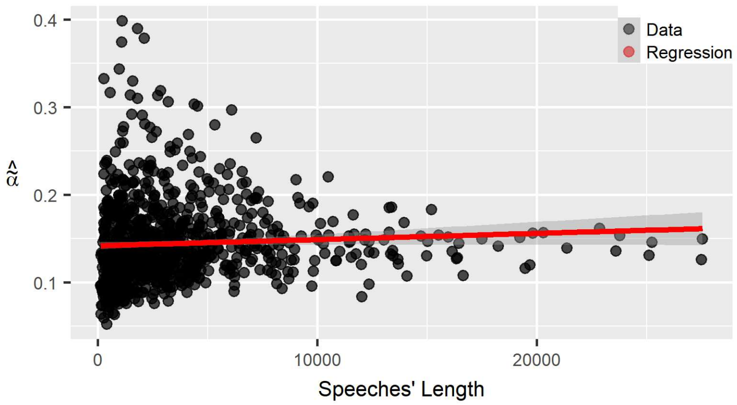

Figure 2.7: Relationship between $\hat{\tilde{\alpha}}$ calibrated for relative frequencies and length of the speeches in term of total number of words used.

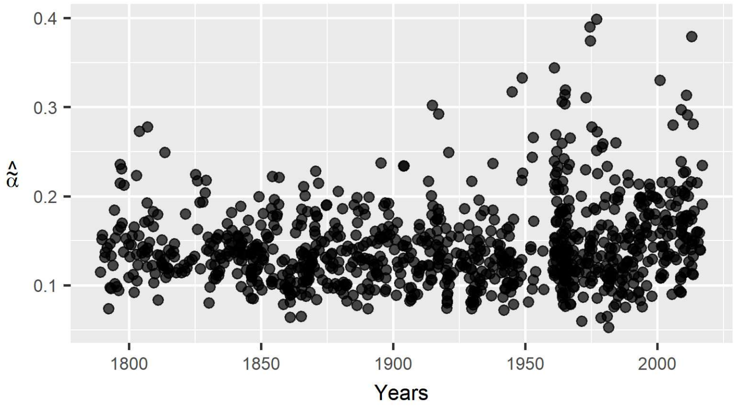

Figure 2.8: Estimated $\hat{\tilde{\alpha}}$ on relative frequencies for each speech over years (eq. (2.2)).

President	No.	President	No.
Lyndon B. Johnson	66	Jimmy Carter	18
Ronald Reagan	57	John Tyler	18
Barack Obama	50	Warren G. Harding	18
Franklin D. Roosevelt	49	Rutherford B. Hayes	16
John F. Kennedy	41	Abraham Lincoln	15
George W. Bush	39	Franklin Pierce	15
Bill Clinton	38	Gerald Ford	14
Woodrow Wilson	33	James Buchanan	14
Ulysses S. Grant	32	William McKinley	14
Andrew Johnson	31	Calvin Coolidge	12
Herbert Hoover	30	William Taft	12
Grover Cleveland	29	Chester A. Arthur	11
Andrew Jackson	26	James Monroe	10
James K. Polk	25	Martin Van Buren	10
Thomas Jefferson	24	John Adams	9
Richard Nixon	23	John Quincy Adams	8
James Madison	22	Millard Fillmore	7
Theodore Roosevelt	22	Dwight D. Eisenhower	6
George Washington	21	Zachary Taylor	4
George H. W. Bush	20	Donald Trump	2
Benjamin Harrison	19	James A. Garfield	1
Harry S. Truman	19	William Harrison	1

Table 2.1: Number of speeches (No.) per President. The list is for 44 presidents, instead of 45, because Grover Cleveland has had two non-consecutive mandates: the first from March 4, 1885 to March 4, 1889 and the second from March 4, 1893 to March 4, 1897. In order to count his number of speeches, the two mandates are grouped together.

Stats	**Speeches Len.**	**Diff. words used**
Max.	27551	3931
Min.	132	76
Median (m)	2315	760
Mean (μ)	3533	916.23
RMS	5256.24	1144.87
St. Dev.(σ)	3893.80	686.85
Variance	15145734.74	471265.19
St. Error	126.27	22.27
Skewness	2.64	1.58
Kurtosis	12	6.02
μ/σ	0.91	1.33
$3(\mu - m)/\sigma$	0.94	0.68

Table 2.2: The numbers in column two offer a statistical summary of the length of the speeches in term of total number of words per speech. The third column contains the statistics of the number of different words used in each speech.

Words	Prob. of fall in the ranks [1-10]
the	9.99%
and	9.97%
of	9.96%
to	9.94%
in	9.38%
a	8.44%
that	6.78%
we	3.33%
our	3.32%
be	3.21%

Table 2.3: The most repeated tokens' probability of having them in the ranks [1-10] of each speech.

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\hat{\alpha}}$	$\hat{\beta}$	$\hat{\gamma}$
Max	4117.92	6.13	1.23	0.40	6.13	1.23
Min	9.72	-0.57	0.54	0.05	-0.57	0.54
Median m	326.35	0.72	0.9	0.14	0.72	0.97
Mean μ	521.13	1.01	0.97	0.14	1.01	0.97
RMS	25.37	0.05	0.03	0.004	0.05	0.03
Standard Deviation	583.69	0.97	0.10	0.05	0.97	0.10
Variance	340339.10	0.94	0.01	0.002	0.94	0.01
Standard Error	18.93	0.03	0.003	0.001	0.03	0.003
Skewness	2.40	1.41	-0.45	1.55	1.41	-0.41
Kurtosis	10.40	5.35	3.65	6.90	5.36	3.53
μ/σ	0.89	1.03	9.53	3.04	1.04	9.53
$3(\mu - m)/\sigma$	-1.00	-0.89	0.20	-0.56	-0.89	0.20

Table 2.4: Statistical summary of the estimated parameters on relative and absolute frequencies in accordance with Eqs. (2.1) and (2.2) respectively.

	R^2	R^2_{rel}
Max	1.00	1.00
Min	0.91	0.91
Median m	0.99	0.99
Mean μ	0.98	0.98
Standard Deviation	0.01	0.01

Table 2.5: Statistical summary of R^2s calculated for each fit with Eqs. (2.1)and (2.2). They represent the models' goodness of fit calibrated over each speeches with absolute and relative frequencies.

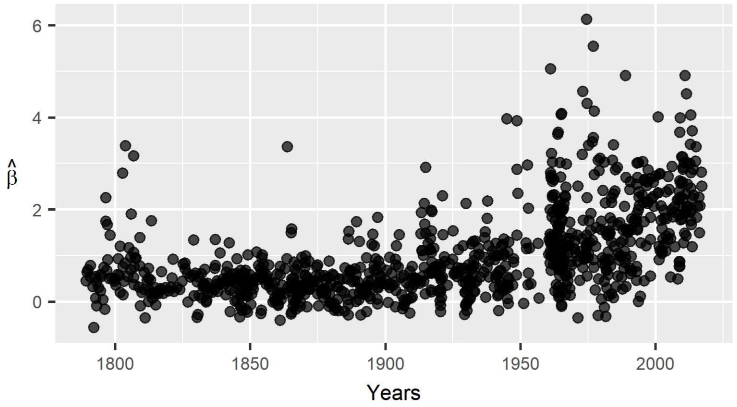

Figure 2.9: Estimated $\hat{\beta}$ on relative frequencies for each speech over years (eq. (2.2)).

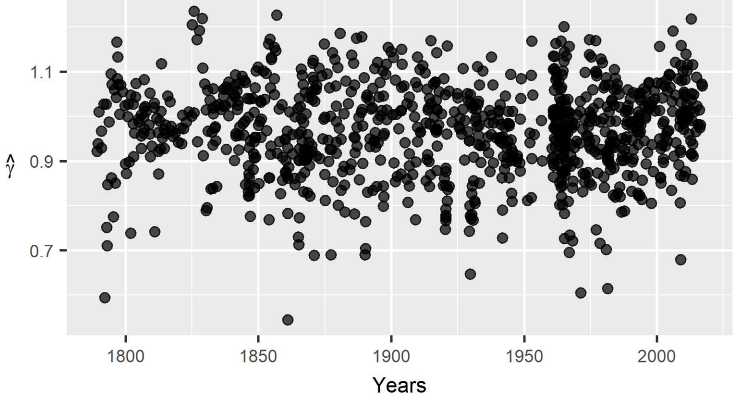

Figure 2.10: Estimated $\hat{\gamma}$ on relative frequencies for each speech over years (eq. (2.2)).

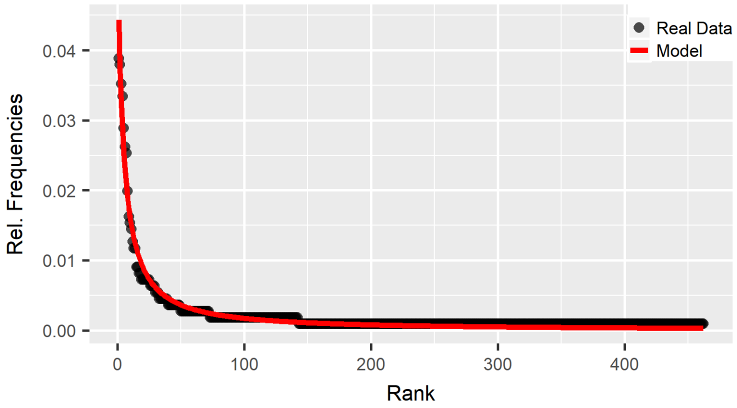

Figure 2.11: January 20, 1977 - Inaugural Address - Jimmy Carter. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the highest $\hat{\hat{\alpha}} = 0.39$; $\hat{\beta} = 5.54$; $\hat{\gamma} = 1.16$; N $=$ 1107; $R^2 = 0.97$.

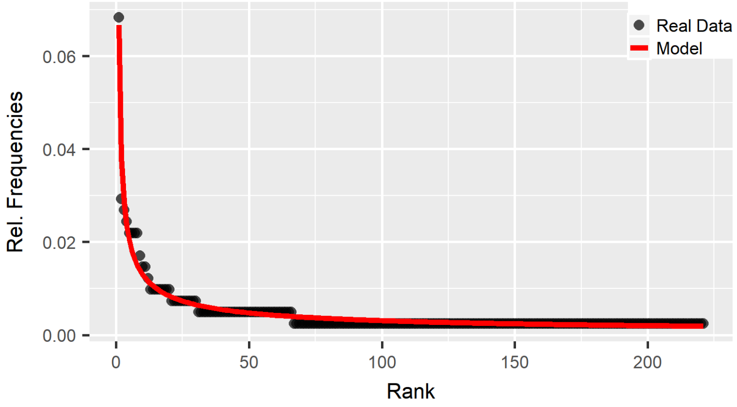

Figure 2.12: August 3, 1981 - Remarks on the Air Traffic Controllers Strike - Ronald Regan. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the lowest $\hat{\hat{\alpha}} = 0.05$; $\hat{\beta} = -0.32$; $\hat{\gamma} = 0.61$; N $= 410$ $R^2 = 0.97$.

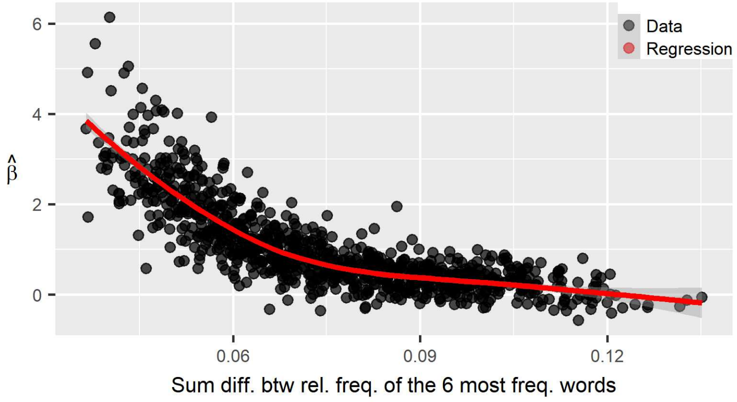

Figure 2.13: $\hat{\beta}$ against the summed differences in relative frequencies of the top 6 repeated words within each speech.

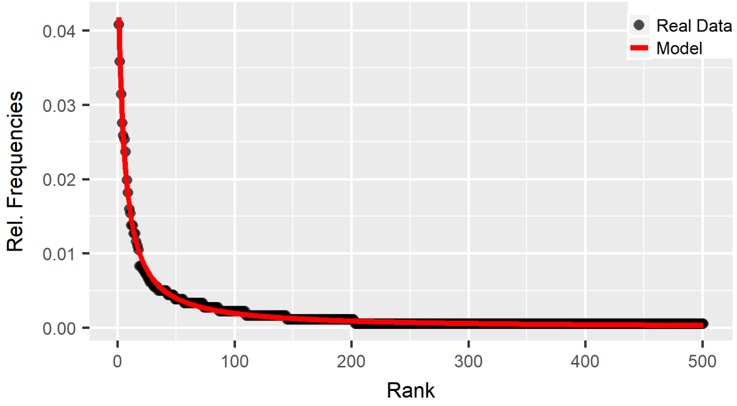

Figure 2.14: August 9, 1974 - Remarks on Departure From the White House - Richard Nixon. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the highest $\hat{\beta} = 6.12$; $\hat{\hat{\alpha}} = 0.38$; $\hat{\gamma} = 1.13$; N = 1815; $R^2 = 0.99$.

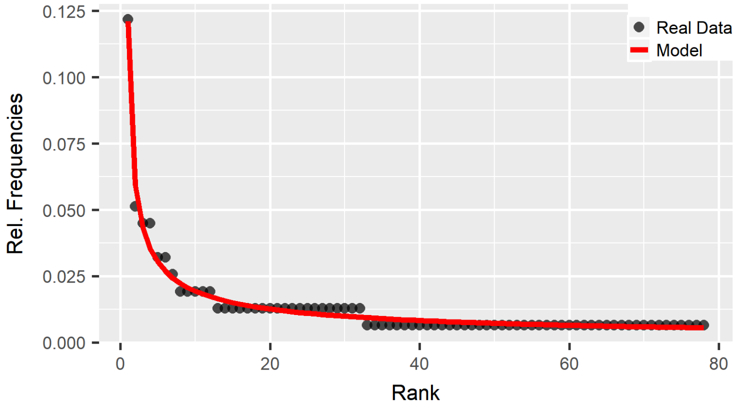

Figure 2.15: April 5, 1792 - Veto Message on Congressional Redistricting - George Washington. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the lowest $\hat{\beta} = -0.56$; $\hat{\tilde{\alpha}} = 0.07$; $\hat{\gamma} = 0.59$; N = 156; $R^2 = 0.98$.

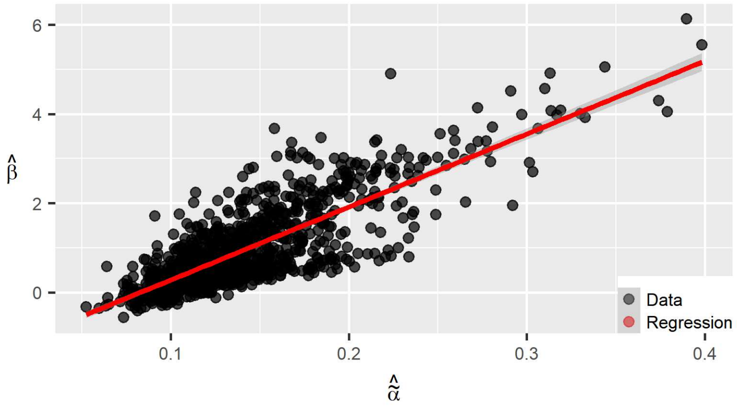

Figure 2.16: Graphical insight of the relationship between $\hat{\tilde{\alpha}}$ and $\hat{\beta}$ in the estimation run using eq. (2.2).

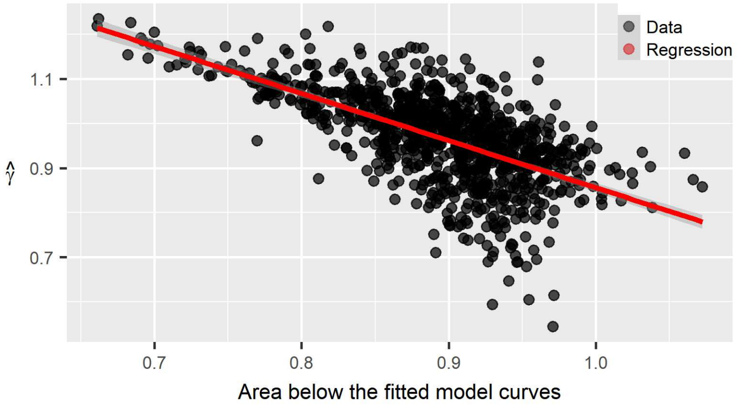

Figure 2.17: The $\hat{\gamma}$ against the areas underlined by each fitted model computed with eq. (2.3).

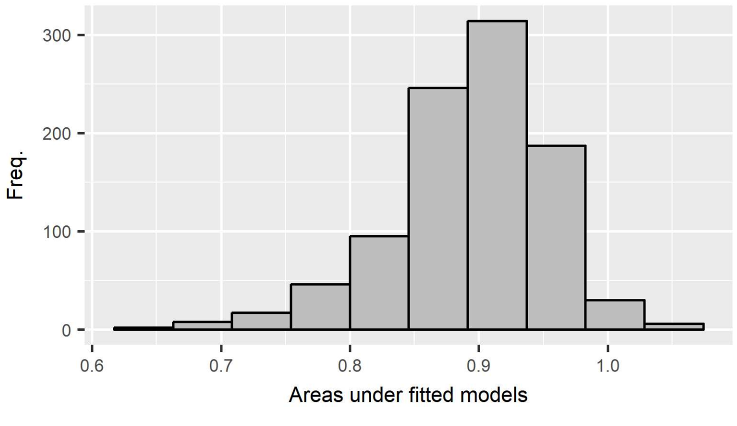

Figure 2.18: Areas under the fitted models computed with eq. (2.3).

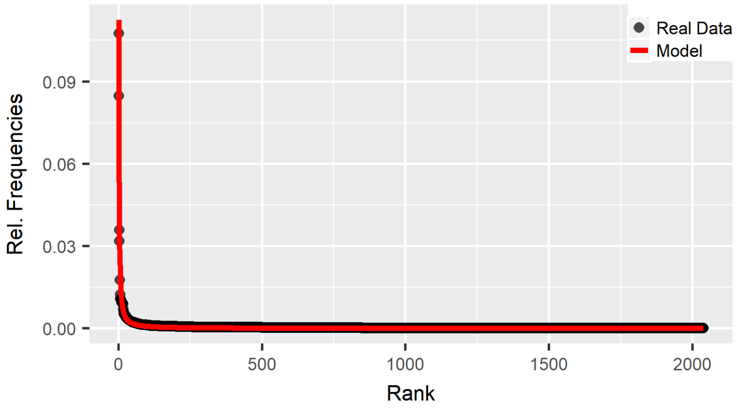

Figure 2.19: December 6, 1825 - First Annual Message - John Quincy Adams. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the highest $\hat{\gamma} = 1.23$; $\hat{\hat{\alpha}} = 0.21$; $\hat{\beta} = 0.70$; N $= 9023$; $R^2 = 0.96$.

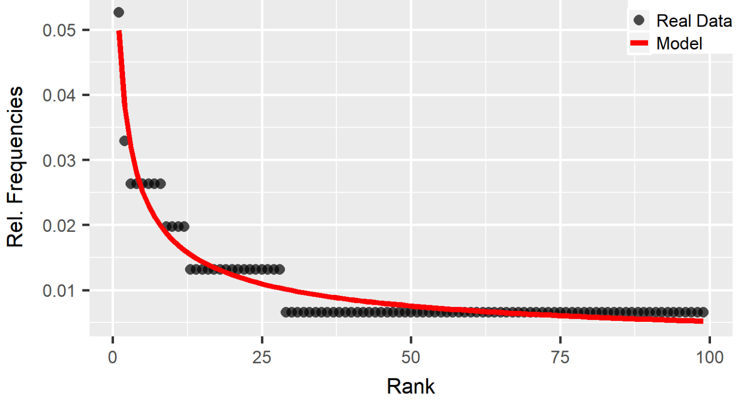

Figure 2.20: "February 11, 1861 - Farewell Address - Abraham Lincoln. Comparisons between real data and fitted models over the speech words relative frequencies for the case of the lowest $\hat{\gamma} = 0.54$; $\hat{\hat{\alpha}} = 0.06$; $\hat{\beta} = 0.58$; N $= 152$; $R^2 = 0.93$.

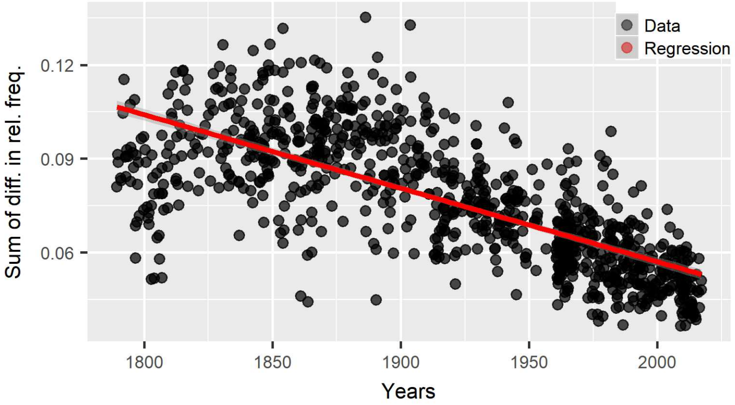

Figure 2.21: Sum of the computed differences between all the words relative frequencies within each speech along the years.

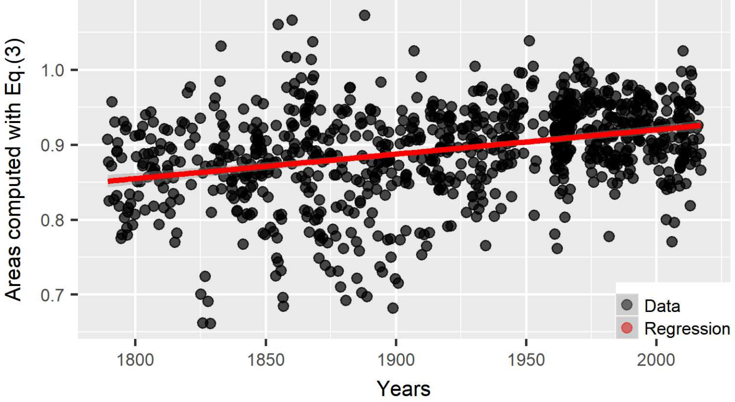

Figure 2.22: Each point represents an area under the respective fitted model computed with eq. (2.3).

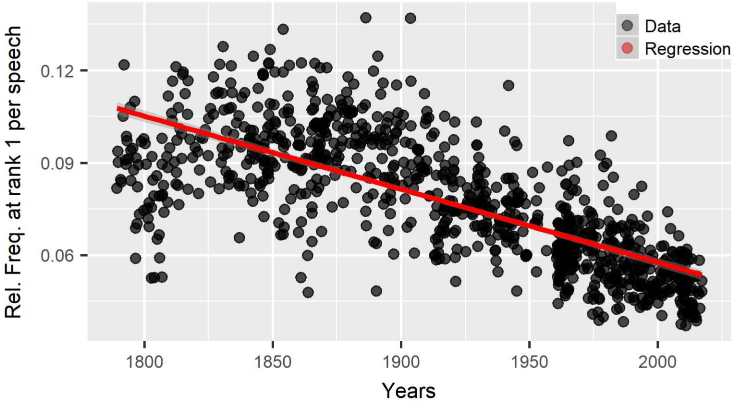

Figure 2.23: The relative frequencies of the most used words in each speech along the years.

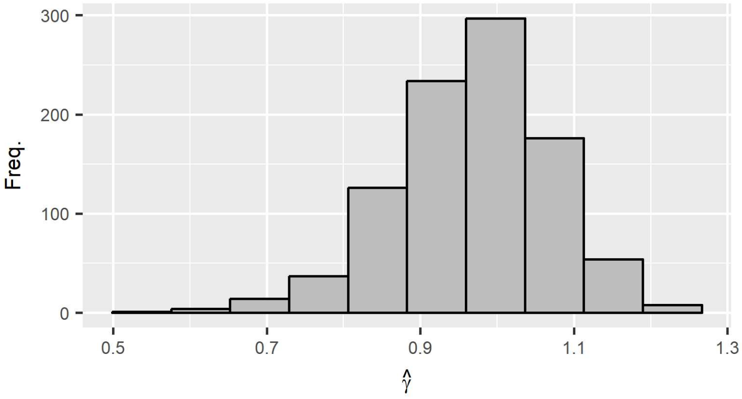

Figure 2.24: Histogram of the $\hat{\gamma}$' values.

Chapter 3

The socio-economic content and implications of political speeches – Part II: relevance of the hapaxes legomenon and of the related core

The official speeches of the US Presidents are, of course, carefully written. Any single locution or term is evaluated, in order to guess what the impact will be on in the audience and in the entire socio-economics environment. This chapter moves from this premise and deals with the analysis of some relevant aspects of the large set of Presidents' speeches presented in Section 2.2. In particular, we study the collection of the hapaxes legomenon in each speech. Indeed, we believe that the decision of saying just once a specific word is carefully planned by the Presidents and their collaborators to meet a precise need, so to deliver a certain message. On the other hand, when such a specific word is pronounced once by several Presidents, then one can guess about a common behavior of such important persons when speaking. Specifically, the message contained in the hapax is recurrent in different circumstances and contexts. Thus to study hapaxes goes much further than usual text content or text structure analysis as in the approach of [108, 35, 127].

The scientific ground of the present study lies in the meaningfulness of the content of the hapaxes. A couple of remarkable examples are worth mentioning.

In the overall work of Giacomo Leopardi, the word *ultrafilosofia* has been used only once for the contextualization of the philosophical system of the author. However, the authoritative Encyclopedia Treccani refers to *ultrafilosofia* to describe the Leopardi's thought. In the related entry, *ultrafilosofia* is no longer a hapax, but it appears 9 times [112].

Mnemosynus is a hapax for the Latin language. In fact, it appears in the entire collection of available writings in Latin only once, in Catullo's Carmina. This term points to the mythological figure of the goddess of memory. Such a hapax is not neglected in subsequent modifications and contamination of the linguistic evolution, and *mnemonic* comes evidently from *Mnemosynus*. Let us mention that hapaxes have been studied elsewhere, like in well known books, as Bible and Quoran (see [6, 133]), which contain speeches attributed to different speakers.

The dataset here considered is the same employed in Chapter 2. The collection of the hapaxes of the speeches is stored in 951 speech-based sets; such sets are merged together. The resulting merged set contains all the words that have been hapaxes at least in one speech, along with their frequencies in the overall set of talks. To clarify this conceptualization of the frequency, think that if the frequency of a hapax is 5, then such a word has been a hapax in five different speeches. This said, the maximum hypothetical frequency of a word is 951 whilst the minimum one is 1.

The study proceeds in three sequential directions.

Firstly, a rank-size relation is assessed over the set of the hapaxes, where *size* is measured through the frequency of the words in the entire set of speeches. In accord with other linguistic studies [98, 110, 8, 9, 50, 15, 118], we test the validity of the Zipf-Mandelbrot law in properly fitting the data [146, 147, 72, 73]. In this preliminary phase, we find statistical compliance of the considered dataset with Zipf-Mandelbrot law, even in presence of (quite) negligible deviations at low ranks (see the third step for a comment on this).

Therefore, after this step, we use the obtained calibrated curve to identify the core of the hapaxes by using the indicator proposed by [12] in the science measurement context of the scientists' coauthors. Such an indicator is a replication of the H-index – where H stands for Hirsch, who invented it in [48] – used to evaluate scientific research. It is crucial to recall that the value of the H-index of a scientist is $\bar{H} \in \mathbb{N}$ when $\bar{H}$ is the maximum number of papers authored by the scientist which have been cited at least $\bar{H}$ times. In this context, the core of the hapaxes is the set with cardinality $\bar{H} \in \mathbb{N}$

which contains the maximum number of hapaxes whose frequency is at least $\bar{H}$. Analogously to the role of the Hirsch index in describing properly the production of a scientist, the core of the hapaxes represents here a meaningful synthesis of the most relevant tokens pronounced by the Presidents. In this respect, the ratio between the area of the core and the one of the entire set of hapaxes – computed with respect to the best-fit curve of the rank-size law – is a percentage measure of the most relevant hapaxes in the overall history of the US Presidents' speeches.

The third step consists of the exploration of the core and of its properties. We show that the core is a set whose hapaxes have ranked frequencies satisfying a Zipf-Mandelbrot law. Furthermore, as already preannounced above, in the present rank-size analysis of the overall set of the hapaxes we have found small deviations at low ranks. This phenomenon has been observed firstly in [55] where the urbanization of cities has been studied. In our specific case, this means that the best fit curve does not represent "perfectly" the scatter plot of the low-ranked hapaxes. The reason for this stands in the outlier-type behavior of a group of mostly used hapaxes. We here guess that such outliers are the hapaxes in the core and redo the best fit procedure by removing the core from the overall original sample. Results confirm the improvement of the fit. According to [61], the token at rank equal to 1 is the so-called *king* whilst the others are the *vice-roys*, and in this case there is a *king plus vice-roy effect*. A similar behavior is shown in Chapter 1. For a further example of this effect, refer to [27]. Later on in this chapter we propose an explanation for this phenomenon in our context.

The rest of the chapter is organized as follows. Section 3.1 contains the illustration of the dataset. Section 3.2 is devoted to the illustration of the methodology used for the analysis. Section 3.3 presents the results and related comments. Last section offers some conclusive remarks.

3.1 Hapaxes legomenon of the speeches

The dataset here employed is a sub-selection of those presented in Section 2.1. Indeed in Chapter 2 we analyze a dataset contains words and their respective frequencies for each talk transcript.

In this study we are interested in tokens occurred just once per speech. Hence we collect from the original dataset all the terms that have frequency equal to one and we store them into a vector. Of course some speeches have hapaxes in common because some words pronounced just once in a talk could be present once in another. For this reason we count their occurrences into the hapax list and we make a table of frequencies where the most present hapax can have a frequency equal to 951, namely it should be a hapax in each talk. From now on, when we refer to frequency, we are speaking about the occurrences into the hapax list. The most common hapaxes are reported in Table 3.1. The maximum frequency is realized by the word *sense*, which appears 250 times as a hapax in a President's speech. It means that there is a term used just once into 250 speeches and other that appear as hapax into one single speech. Indeed, there is a list of 10088 tokens which are hapaxes only in one speech (thus, having unitary frequency in the hapax list). The whole resulting table is made by 31074 hapaxes. The principal statistical indicators can be found in Table 3.2, and it gives an idea of the frequencies statistical features.

Despite of the manipulation and correction processes documented in Section 2.1, some minor typos are still reported into the dataset of hapaxes. But they do not exceed the 2%. A visual inspection of them allows to conclude that the majority falls into the terms that occur just once into the hapaxes list. Therefore they leads to a negligible effect on the analysis object of the present study.

3.2 Methodology

The matrix that contains the hapaxes is ranked in decreasing order according to the frequencies of the terms. In this respect, the size of a word is its frequency of occurrence into the hapax list. In the rank-size analysis, we will denote size and rank by s and r, respectively.

The Zipf-Mandelbrot law is used for fit, according to the following rule:

$$f(r) = \frac{\alpha}{(\beta + r)^{\gamma}}, \tag{3.1}$$

where α, β, γ are parameters to be calibrated for fitting the sample under investigation.

As we will see, there is a very good compliance of the considered data with the Zipf-Mandelbrot law (see Table 3.3 and Figure 3.1 in Section 3.3). Such a property can be used to define the measure of the core of the hapaxes.

In fact, the core of the hapaxes is defined through the H index, in a similar way in which it has been introduced by [48] to evaluate scientific

Word(s)	**Frequency**
sense	250
given	247
bring, house	240
give	239
hand, themselves	229
within	228
others, therefore	225
set	224
take	222
second	221
find, full, making, since	220
among	217
again, does,	215
itself, remain	214
being, brought, done, soon, whose	213
part, protect	212
known, small	211
able, beyond, carry, friends	210
call, day, far, fellow, means, opportunity, then, washington, while	209
course, order, single	208
essential, important, meet, reason	207
another, left, like, respect, seen	206
certain, few, necessary, possible, purpose	205

Table 3.1: The most frequent 61 words, along with their frequencies.

Statistical indicator	**Value**
Number of data	31074
Minimum	1
Maximum	250
Mean (μ)	16.38
Median (m)	3
RMS	36.09
Standard Deviation (σ)	32.16
Variance	1034.30
Standard Error	0.18
Skewness	3.24
Kurtosis	11.60
μ/σ	0.51
$3(\mu - m)/\sigma$	1.25

Table 3.2: Main statistical indicatorsof the hapaxes found in the speeches .

research. Specifically, such an index is $\bar{H}$ when $\bar{H}$ is the maximum number of words whose frequency is at least $\bar{H}$. The resulting set of $\bar{H}$ words is the core of the hapaxes.

By employing $\bar{H}$ and the best-fit curve defined in eq. (3.1) and with parameters in Table 3.3, we are able to provide an absolute and relative measure of the core of the hapaxes. We denote such measures as $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{R}}$, respectively. They are defined as the area of the region below the curve in eq. (3.1) delimited by $r = 1$ and $r = \bar{H}$ and as the ratio between such area and the area of the overall region, from $r = 1$ to $r = 31074$, respectively. Specifically, the absolute measure of the core of the hapaxes is

$$\mathcal{M}_A = \int_1^{\bar{H}} \frac{\hat{\alpha}}{(\hat{\beta} + r)^{\hat{\gamma}}} dr, \tag{3.2}$$

while the relative measure is

$$\mathcal{M}_R = \frac{\mathcal{M}_{\mathcal{A}}}{\int_1^{31074} \frac{\hat{\alpha}}{(\hat{\beta}+r)^{\hat{\gamma}}} dr}. \tag{3.3}$$

$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
6.029×10^8	2540	1.896

Table 3.3: Best-fit parameters of the Zipf-Mandelbrot law with eq. (3.1).

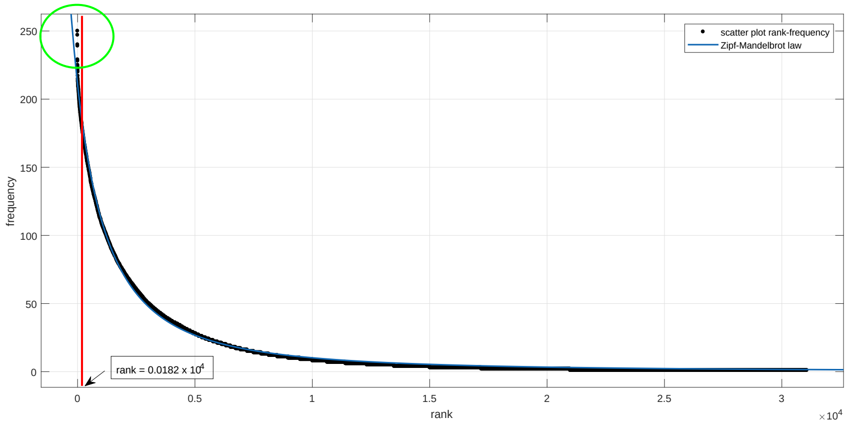

Figure 3.1: Best-fit curve, according to equation (3.1) and calibrated parameters in Table 3.3. The scatter plot of the original sample is juxtaposed for a better comparison; the agreement is very good, data and fits are hardly distinguishable from each other. Notice the slight deviations at low ranks (green circle in the figure), suggesting to the presence of king and vice-roy effects (see [61]). The red vertical line points to $\bar{H} = 182$, which delimitates the core of the hapaxes.

3.3 Results and discussion

The results of the best-fit exercise are reported in Table 3.3, where one can find the calibrated parameters. The value of R^2 is 0.9971, which suggests a quite perfect compliance of the considered ranked dataset with the rank-size Zipf-Mandelbrot law. Figure 3.1 further supports such a result by proposing a visual inspection of the fit.

By looking at the data we have that $\bar{H}$ is 182, i.e. there exist 182 words whose frequency is at least 182 and, simultaneously, there are not 183 words with frequency at least 183. The most frequent hapaxes are reported in Table 3.1 for the reader convenience. To save space, only one third of the core is shown, i.e. the most frequent 61 hapaxes. Thus, by applying formulas (3.2) and (3.3) and by using the values listed in Table 3.3, a straightforward computation gives that

$$\mathcal{M}_A = \frac{\hat{\alpha}}{-\hat{\gamma}+1}\left[(\hat{\beta}+\bar{H})^{-\hat{\gamma}+1} - (\hat{\beta}+1)^{-\hat{\gamma}+1}\right] = 35783.9769 \qquad (3.4)$$

and

$$\mathcal{M}_R = \frac{(\hat{\beta}+\bar{H})^{-\hat{\gamma}+1} - (\hat{\beta}+1)^{-\hat{\gamma}+1}}{(\hat{\beta}+31074)^{-\hat{\gamma}+1} - (\hat{\beta}+1)^{-\hat{\gamma}+1}} = 0.0664 \qquad (3.5)$$

Notice that the hapaxes contained in the core represents a negligible percentage – about 0.58% – of the entire set of words said once. However, in terms of frequencies, we have that the core is 6.64% of the overall set, as the relative measure assures. This means that a very small set of words have been selected to be said only once in a large number of speeches, with about eleven times the frequencies over the hapaxes.

To have a view of the set of the core, we report in Table 3.4 the main statistical indicators for the frequencies of the set of such 182 hapaxes.

By exploring the core of the hapaxes itself, one can see that the frequencies of the tokens therein contained are well fitted by a Zipf-Mandelbrot law. Refer to Figure 3.2 and Table 3.5 for the details. The statistical goodness of fit is rather satisfactory also in this case, with $R^2 = 0.978$. A visual inspection confirms the very good compliance of the data with Zipf-Mandelbrot law even if there are some evident deviations (see Figure 3.2) with respect to the case of Figure 3.1.

Notice that the size of the sample is able to affect the goodness of fit – in some circumstances, one can claim that larger size leads to less scattered data – hence letting our result coherent with our expectations.

Statistical indicator	**Value**
Number of data	182
Minimum	183
Maximum	250
Mean (μ)	199.65
Median (m)	197
RMS	199.64
Standard Deviation (σ)	13.54
Variance	183.27
Standard Error	1
Skewness	1.12
Kurtosis	1.46
μ/σ	14.74
$3(\mu - m)/\sigma$	0.59

Table 3.4: Main statistical indicators related to the set of the hapaxes in the core.

$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
287.7	5.903	0.084

Table 3.5: Best-fit parameters of the Zipf-Mandelbrot law with eq. (3.1) for the case of the hapaxes belonging to the core.

The hapaxes in the core produce a king (the word *sense*, with frequency 250) and 181 vice-roys effect. Indeed, once the core is removed from the sample, one obtains a perfect fit through a calibrated Zipf-Mandelbrot law, with the removal of the deviations at the low ranks (compare Figures 3.1 and 3.3). In the case of core removal, the goodness of fit parameter remains quite perfect, with $R^2 = 0.9965$. The best fit parameters can be found in Table 3.6.

3.4 Conclusions

This chapter completes the challenging theme of exploring a part of the structure of the US Presidents' speeches. We start from the premise that speeches are official and carefully written. Each of them has some messages

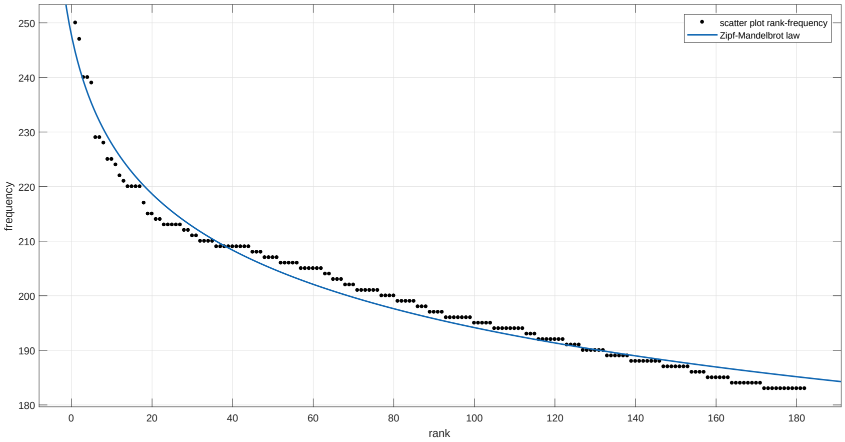

Figure 3.2: Best-fit curve, according to equation (3.1) and calibrated parameters in Table 3.5 for the case of the hapaxes in the core. The scatter plot of the original sample of the core is also shown for comparison purposes; the agreement is visually good.

$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
4.359×10^8	2668	1.861

Table 3.6: Best-fit parameters of the Zipf-Mandelbrot law with eq. (3.1) for the case of all the hapaxes without those belonging to the core.

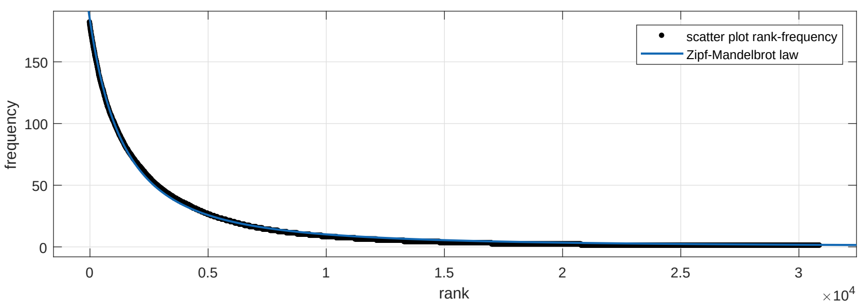

Figure 3.3: Best-fit curve, according to equation (3.1) and calibrated parameters in Table 3.6 for the case of the hapaxes excluding the core. The scatter plot and the fitted curve are not distinguishable. The deviations at the low rank shown in Figure 3.1 do not appear, thus leading to the statement of the presence of king and vice-roys effects for the elements of the core in the respect of the overall sample.

to deliver to different audiences and to the entire society. Thus, words are tactically selected with care.

We are here interested in the hapaxes of each speech. For us they represent a relevant aspect of the Presidents communication strategy. In fact, one can observe some recurrent hapaxes, which are consciously decided to be pronounced only once in several occasions and by several Presidents. We assume that the relative rarity of these words is thought to be intentional, sometimes appearing as new (or astute) terms, implying the President modernity, elitism, and wide knowledge. So the use of the hapaxes is important to have a deep understanding of what there is behind the talks.

If appropriately merged and ranked, hapaxes show regular paths and can be successfully fitted by Zipf-Mandelbrot law. Moreover, there is a privileged set of core hapaxes, defined through the introduction of a Hirsch-based threshold.

We show that a very small number of words have been pronounced once several times in official speeches. This lets us understand the presence of common messages and arguments in the historical paramount view of the US Presidents' interventions. The set of such words represents the core of the hapaxes. Such a core can be interpreted as those words which strike a point, even though they are rarely used.

We also show that the core has a structure similar to the one of the overall sample, with a compliance with a rank-size law of Zipf-Mandelbrot type.

Moreover, the core is also responsible for deviations of the overall set of hapaxes from the best Zipf-Mandelbrot curve. In this, king and vice-roys effects are detected.

The present analysis represents a further step towards the comprehension of the changes of the political speeches and how messages are – and have been – delivered by the US Presidents.

Chapter 4

The socio–economic content and implications of political speeches – Part III: effects on the S&P 500 index

The US President is one of the most influential person in the world. His decisions have an high impact on many aspects of the citizens' life not only in US but also in other countries (see e.g. [78, 129] for considering some effects of political decisions). Consequently, the US President's actions and the respective announcements are constantly monitored by the news agencies and they constitute a relevant field of research (see [134, 80] as examples of political communication studies).

In this chapter, we want to check if and how much the US stock market is affected by the Presidents' public communications. Specifically, the scope of this chapter is to explore the impact of the US Presidential talks related to economics on the Standard and Poor's 500 (S&P 500 hereafter).

In order to do it we use two datasets. The one presented in Chapter 2 that covers the entire US history spanning from April 1789, with George Washington presidency, to February 2017 with the last President elected, Donald Trump; the second one is given by the daily prices, volumes and returns of the S&P 500 index. The source of the former dataset is the Miller Center `https://millercenter.org/`, a prestigious Political Research Institution affiliated to the University of Virginia; in line with a wide set of authoritative financial contributions (see e.g. [7, 57]), we take as a source of the latter dataset the freely available website "Yahoo! Finance" https://it.finance.yahoo.com. The considered period for the S&P 500 data starts – on the basis of data availability – at March 1^{st}, 1950 and ends at March 2^{nd}, 2017. The final time is selected by consistency with the speeches dataset, which ends two days before with the lastly considered Trump's talk.

The economic content of the speeches is quantified through the assessment of the presence of terms whose meaning is attributable to the semantic area of economics. In doing so, we follow the steps of [17], where the authors determine a class of terms ascribable to economic uncertainty. Differently from them we use the glossary presented in [21] which is particularly appropriate, being also employed by *The Economist* as a source of locutions belonging to the economics glossary. To further improve the reliability of such data, the aforementioned list is expanded by including the terms listed in the Wikipedia's glossary of economics (https://en.wikipedia.org/wiki/Glossary_of_economics).

This text analysis approach has been used in several context and for an high number of phenomena explanations. We mention [42, 7], [131] for the specific case of the sentiment analysis. In [1] the authors have counted the occurrences of pre-classified characters combinations. Namely the emoticons used in a large set of tweets are taken into account along with their positive and negative meanings, in order to gain insights about the text sentiment. It is also worthy to mention [25], where the authors have considered corporate governance news and its interaction with the corporate economic and financial standing. Their results show that the market actors do not appreciate the news about firms ownership for the case of profitable companies. Furthermore, they found some traders able to anticipate the news about corporate governance, thanks to the rumors on that topic. Finally they have shown that the news release affects the traders behaviors. Their study is developed by using Wordsmith 4 (see [122]), a computer program used to count the occurrences of class of words into a texts.

In [37] the authors have made an economics blogs pessimism indicator using a list of words from the Harvard IV dictionary and the Lasswell value dictionary. After this preliminary step, the authors designed some investment strategy based on the considered indicators. We also mention [47], which contains an extensive investigation of the institutional tradings in presence of news; in this study, it clearly emerges that the institutions' financial positions are influenced by the news even before information becomes public.

In our study, we treat the speeches as if they are news for the stock market. In particular, we compare the economic terms of the speeches with S&P 500 by considering three stances: S&P 500 financial variables at the date of the speech, one day ahead and one day before. In particular, when the talks have been delivered at open market, then S&P 500 prices, returns and volumes are available at the same date. This is not true in the cases in which the Presidents have spoken during the week ends or during holidays. In Table 4.1 we present to the reader how many speeches have been delivered for each day of the week within our dataset.

From a methodological point of view, the investigation of the effects of economics related talks on the S&P 500's index is implemented under two different perspectives. By one hand, a statistical physics approach is employed. Precisely, we use the Kendall's τ correlation, introduced by [59], for measuring the ordinal association between the frequencies of the economic terms in the speeches and the considered financial variables. In so doing, we are able to assess the presence of a relation at a rank level between the occurrences of economic terms in presidential talks and the value of daily prices, returns and volumes of the considered index (for further information about rank-rank correlation see [82, 43]). By the other hand, as a side scientific product, we apply a geometric and information theory point of view by introducing and discussing the distance between the time series of the economic terms frequencies (relative and absolute) and the S&P 500's variables listed above. Specifically, 4 distances are computed: max, min, Manhattan and Euclidean distance. Such distances provide different information on the relationship between the economics glossary of the Presidents and the evolution of the S&P 500. Furthermore, we also calculate the Shannon entropy ([123]) to obtain some information on distributions of each series.

Several studies explore the rank-rank correlations in different *contexts. For example* in [14] there is comparison of UEFA and FIFA rankings of countries' soccer teams, in [45] this correlation measure is applied in the context of archaeology, in [145, 12, 126] there are interesting applications of it for facing some scientometrics problems, in [26] there is an example of Kendall's τ coefficient application in economics geography and [32] provides a clearer empirical results of dependence between two serial variables: reputation and individual rankings of faculties.

The rank-rank correlation approach has been often employed in the text analysis field (see [18, 130, 105]). Furthermore, it has been also used in the sentiment analysis studies as in [137] or [107]. Differently from the quoted papers, we want to highlight the influence of a certain class of words when they are pronounced by US Presidents. To the best of our knowledge, the Kendall's correlation has never been used to investigate the relationship of the economic content of the political messages with financial markets. Furthermore, the numbers of papers that highlight a correlation between stock market and political speeches by using text mining techniques and rank-rank correlation is small. A study particularly close to our approach, even if it is not about political messages analysis, is [114], where the authors used the Kendall's τ to measure the relationship of Google query volumes of search terms related to finance with the stock market.

Distance measures are becoming more and more relevant, especially in the context of machine learning algorithms (see [102] for a wide review of machine learning application for stock market prediction due to text mining methods). Specifically, in the field of news classification and news impact studies, the distances are a fundamental pillar. For example in [40] Euclidean distance is used to find the proper segmentation of some stock time series and then to compare them with different classes of news grouped by key words collections, or [64] where the Euclidean distance is used to determine stock market trends clusters. For a general overview of distance measures between time series see ([31]).

For the peculiar case of text analysis, distances are implicitly employed in studies where trading and prediction algorithms are developed on the bases of text mining technique for news explorations like [97, 125, 106, 142, 121].

Our approach is different from the quoted papers, because we compute distances between normalized time series of different type of data (economic terms and S&P 500) in order to have an idea of the co-movements of the variables of interest. So, even if we use the distance measures in a different way, our scope is similar to those papers that investigates the news impact because we want to question the influences of economic terms' frequencies on S&P 500.

Lastly, the Shannon entropy is often used in the context of text mining, especially into text categorization problems as in [62], in authorship attribution problems (see [117]) and in the financial sector to analyze some financial time series features ([19]). In our case we use the Shannon entropy to compare the time series disorder because we are interested in finding common distributional behavior between them.

The results obtained are properly compared with the information grasped from the application of the other approaches.

The chapter is organized as follows: in the next section we present the details of the dataset along with their main statistical features. In Section 4.3 we describe the methodological toolkit. Sections 4.4 and 4.5 contains the results and the conclusion. Section 4.6 reports some conclusive remarks.

Day of the week	No. of speeches delivered
Tuesday	240
Monday	196
Thursday	151
Wednesday	151
Friday	98
Saturday	75
Sunday	40

Table 4.1: Number of speeches delivered for each day of the week in the period between 30/04/1789 and 28/02/2017

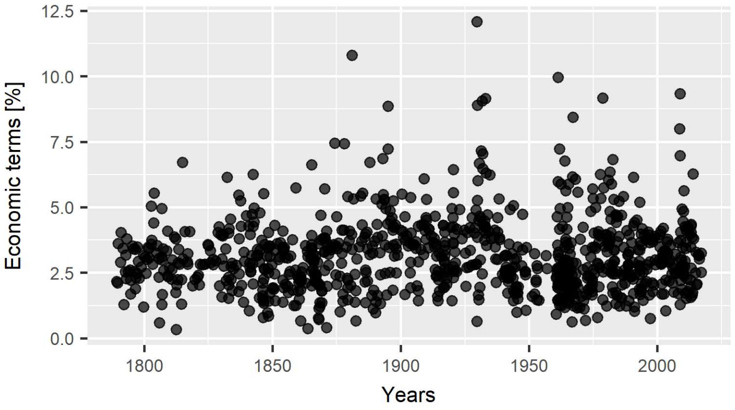

Figure 4.1: Percentage of economic terms occurrences per each speech along the years

4.1 Employed Data

The Standard and Poor 500 is chosen as the most representative index of the US stock market. The daily data of the index closing prices and volumes is downloaded from the "Yahoo! Finance" web site where it is available from

January 3^{rd},1950. The index history is considered until March 2^{nd}, 2017, because the last speech present in the list described in Section 2.2 dates back to the February 28^{th}, 2017 and a couple of days are added to that date for considering the effects on the days after the speech. The daily returns are computed in order to explore the effects of the Presidents' talks on them.

The economic terms are manually taken from [21] and for having a wider economics glossary, the list of strings presented in `https://en.wikipedia.org/wiki/Glossary_of_economics` are added.
On the raw list of locutions is applied a pre-process phase detailed in the next section. Finally, the economics glossary is made by 1483 locutions and the economic content index of the speeches is based on the frequencies of occurrences of them.

The transcripts of the speeches used in this chapter are described in Sections 2.2. A bunch of 951 talks are considered without tokenizing them as in step 10 of Section 2.1 even if all the others phases still remain applied. Indeed, for the purpose here pursued, each whole transcript is needed for counting the glossary terms occurrences into each talk. Since some locutions have more than one single words, if the speeches are tokenized, the research of these terms would be much more difficult.

4.2 Data Mining

In this section is described the process implemented for making the dataset ready to be analyzed.

Rough data on the speeches is properly treated in a devoted processing phase. In particular, the occurrences of all the economic terms in each speech are opportunely stored.

We now enter the details, and describe the process implemented in order to make the data ready to be analyzed.

First of all, a visual inspection leads to the awareness of the presence into the economic terms list of first names and surnames of celebrated economists like John Maynard Keynes, Friedrich Hayek, Karl Marx and Milton Friedman. Such strings are reasonably replaced with the surnames, since it is plausible that a President mentions only the surname when referring to an economist. Therefore, such a replacement does not change the frequencies and avoids unappropriate removals of locutions. Secondly, singular and plural forms of the same terms are considered as a unique word. Finally, some locutions present in both The Economist and Wikipedia's glossary are taken just once. So the number of elements into the resultant economics glossary is 1483.

Once the list of locutions is refined, the sum of the economic terms occurrences into each speech is computed.

The result is an integer number for each speech. Such a number represents the absolute frequency of the words belonging to the considered economics glossary, and it is given at a talk level. The relative frequencies of the economic terms are therefore computed by dividing the mentioned integers by the length of the speeches.

We acknowledge the presence of a small bias in the definition of the relative frequencies. Indeed, some of the locutions belonging to the economics glossary are composed by more than one word – as an example, 'unemployment trap' – but they are considered as single economic words. Therefore, the relative frequency of all the economic words – which is obtained by dividing the economic words aggregate frequency by the speech length – might not have one as an upper bound. However, the effect of such a bias is quite negligible and does not undermine the goodness of the analysis.

Figure 4.2 exhibits the relative frequency of all the economic terms into each speech along the years.

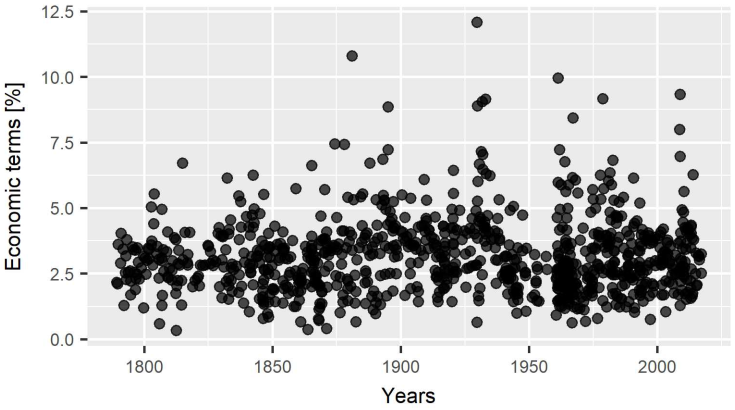

Figure 4.2: Relative frequency of all the economic terms per each speech along the years

For comparison purposes, all the data on absolute frequencies of the speeches and on S&P 500 daily closing prices, volumes and returns are normalized. In particular, for any t – representing a day for the financial variables and a speech for the US Presidents' talks list – the normalized datum is

$$x'_t = \frac{x_t - min(x)}{max(x) - min(x)} \tag{4.1}$$

where x_t is the corresponding original series value, whilst $max(x)$ and $min(x)$ are the intuitive writing of the maximum and minimum over the elements of the entire series. We stress that the relative frequencies of economic terms are in $(0, 1)$ by definition, and a normalization phase is not needed for them.

The last step of the dataset building procedure consists of the truncation of the number of speeches. Indeed, to compare the speeches with the S&P 500 variables, all the talks delivered before January 3^{rd}, 1950 are not considered. In this way the residual transcripts and the respective economic content indicator values are 380.

4.3 Methodologies

To explore the relationship between the speeches economic terms and the S&P 500 index, several cases are considered. For an easy reference in the section of the results, we list them. In particular, the couples under analysis are the following:

(a) S&P 500's normalized returns observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the relative frequencies of the considered economics glossary.

(b) S&P 500's normalized returns observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the normalized absolute frequencies of the considered economics glossary.

(c) S&P 500's normalized closing prices observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the relative frequencies of the considered economics glossary.

(d) S&P 500's normalized closing prices observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the normalized absolute frequencies of the considered economics glossary.

(e) S&P 500's normalized volumes observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the relative frequencies of the considered economics glossary.

(f) S&P 500's normalized volumes observed the same days of the President's talks, the day after and the day before. Each of them is coupled with the normalized absolute frequencies of the considered economics glossary.

The analysis of the day after and before has the relevant meaning of understanding if there are anticipatory effects of the speeches on the market or, conversely, whether the market seems to anticipate the economics themes of the speeches.

Notice that the availability of the financial data depends on the day of the speech. As an example, a speech delivered during the weekend cannot be compared with the "at the date" variables. For consistency, such cases are removed from the list of analyzed speeches when the "at the date" analysis are performed . In details, 41 talks have been stated during the weekends, hence they do not have respective values for the "at the date" S&P 500. Furthermore, for the analysis of contemporaneous effects of the speeches on the S&P 500 index, 8 other talks are excluded because in their respective dates the closing prices, volumes and returns are not available. After this phase, there are 331 speeches to analyze. To have an idea of the talks distribution, Table 4.1 shows the days of the week of the speeches for the entire dataset (see Chapter 2 for further references on the bunch of talks here in analysis).

The correlation analysis is run through the Kendall's τ rank-rank correlation. Such an indicator is computed for couples of series which are jointly observed and with the same cardinality – say (k_i, h_i) with $i = 1, \dots, N$. The computation procedure starts from ranking in increasing (or decreasing) order separately the k's and the h's. Then, the ranks are coupled on the basis of the original joint observations. Kendall rank correlation coefficient τ can be defined as:

$$\tau = \frac{(\text{no. of concordant pairs}) - (\text{no. of discordant pairs})}{N(N-1)/2} \tag{4.2}$$

The number of concordant pairs is given by the number of couples (k_i, h_i) and (k_j, h_j) for which the ranks for both agree after that they have been sorted. Consequently, the number of discordant pairs is given by the couples with discordant ranks.

This measure gives information about the rank-rank correlation between two variables. It allows to conclude about common regularities between different variables equally sized, and it is also called measure of association. For a wide review of the Kendall's τ and other similar measures, see [60].

To compute the Kendall's τ we use the R package: *Kendall* presented in [79]. The Kendall's coefficient is calculated on the couples of array described in (a)–(f) by sorting one of the two element of the couples, therefore influencing the ranks of the remaining element of the couples in accord with the sorting order decided (descendent or crescent). In this way, the Kendall's τ takes into consideration the number of concordant and non-concordant pairs, in order to quantify the singularities in a probable relationship (for further information about rank-rank correlation see [82]). The use of a rank-rank correlation is coherent with the common thread considered because the present work is mainly devoted to a data analysis through a rank-size approach.

Other explorations are performed as a side scientifc research over the datasets for having a more clear view of the comparison between speeches and the S&P 500 variables.

In particular, some distance measures are computed between the economics glossary terms (normalized absolute and relative) frequencies and the S&P 500 normalized data. The analyzed cases are (a)–(f) of the list above.

The distance measures selected are:

$$d_{mx}(f, S\&P500^{(k)}) = max_t|f_t - S\&P500_{t+k}| \tag{4.3}$$

$$d_{mn}(f, S\&P500^{(k)}) = min_t|f_t - S\&P500_{t+k}| \tag{4.4}$$

$$d_{am}(f, S\&P500^{(k)}) = \frac{1}{T^{(k)}}\sum_{t=1}^{T^{(k)}}|f_t - S\&P500_{t+k}| \tag{4.5}$$

$$d_{ec}(f, S\&P500^{(k)}) = \sqrt{\sum_{t=1}^{T^{(k)}}(f_t - S\&P500_{t+k})^2} \tag{4.6}$$

where $k = -1, 0, 1$; $S\&P500^{(k)}$ is the general indicator of the normalized S&P 500 data (volumes, closing prices or returns) observed one day before, at the day and one day ahead for $k = -1, 0, 1$, respectively; $T^{(k)}$ is the total number of observations that varies with respect to the time selection; f_t is the summed frequency (absolute normalized or relative) of the economic terms at time t.

Furthermore, the Shannon entropy (see [124]) for each series of data is computed and compared thanks to the R package: *entropy* (see [46] for further information). In so doing, we quantify the information contained by the series and discuss their closeness. At this aim, the variation range of each series is divided in N intervals of equal size.

Thus, entropy reads as

$$H = -\sum_{j=1}^{N} p_j \log_2 p_j \tag{4.7}$$

where p_j is the probability of having an observation in the class j, and it is empirically determined by the frequencies of the original sample. In our experiment, we have tried several values of N and finally decided to set $N = 280$.

4.4 Results

Tables 4.2, 4.3, 4.4 contain the statistical summaries of the economic terms' relative and absolute normalized frequencies and the S&P 500 normalized variables for each case respectively (contemporaneous and the two out-of-phase adjustments).

The Kendall's τ coefficients calculated for measuring the rank-rank correlations of the pairs (a),(b),(c),(d),(e) and (f) are reported in Table 4.5.

Distances calculated in accord to formulas (4.3), (4.4), (4.5) and (4.6) are shown in Table 4.6. Furthermore, Shannon entropy as in formula (4.7) is presented in Table 4.7.

4.5 Discussion of the results

Figure 4.3 presents the histograms of the economic terms' relative frequencies. They are asymmetric, with positive skewness which suggests a right tailed distribution (see also Tables 4.2, 4.3 and 4.4). Furthermore the value of the kurtosis indicates a leptokurtic behavior. There is an evident presence of outliers: for example, a speech exhibits 7.5% of economic terms.

The histograms of normalized absolute frequencies are shown in Figure 4.4. The asymmetry indexes in Tables 4.2, 4.3 and 4.4 manifest the presence of skewed distributions, with the kurtosis that are almost doubled with

	Economic	Economic	Prices Norm.	Volume Norm.	Ret. Norm.
	Rel. Freq.	Abs. Freq. Norm.			
N. Obs.	331	331	331	331	331
Max	0.075	0.862	0.986	0.685	0.874
Min	0.003	0.001	0.002	0.000	0.516
Median m	0.018	0.067	0.049	0.005	0.678
Mean μ	0.021	0.104	0.190	0.083	0.680
RMS	0.024	0.147	0.295	0.173	0.681
St. Dev. σ	0.011	0.104	0.226	0.151	0.032
Var.	0.000	0.011	0.051	0.023	0.001
Sd. Err.	0.001	0.006	0.012	0.008	0.002
Skewness	1.542	2.228	1.264	1.944	0.564
Kurtosis	6.371	11.967	3.562	5.804	9.708
μ/σ	1.847	1.004	0.841	0.551	21.179
$3(\mu - m)/\sigma$	0.686	1.081	1.867	1.552	0.191

Table 4.2: Statistical summary of the variables used for evaluating the impact of the speeches on the S&P 500 observations on the same days in which the talks are delivered.

	Economic	Economic	Prices Norm.	Volume Norm.	Ret. Norm.
	Rel. Freq.	Abs. Freq. Norm.			
N. Obs.	377	377	377	377	377
Max	0.075	0.862	0.989	0.651	0.799
Min	0.003	0.003	0.002	0.000	0.519
Median m	0.018	0.061	0.049	0.005	0.677
Mean μ	0.020	0.099	0.188	0.077	0.677
RMS	0.023	0.142	0.294	0.160	0.678
St. Dev. σ	0.011	0.102	0.227	0.141	0.029
Var.	0.000	0.010	0.051	0.020	0.001
Sd. Err.	0.001	0.005	0.012	0.007	0.002
Skewness	1.615	2.390	1.291	1.935	-0.255
Kurtosis	6.767	12.628	3.619	5.704	8.849
μ/σ	1.853	0.963	0.829	0.550	23.121
$3(\mu - m)/\sigma$	0.686	1.105	1.840	1.546	0.001

Table 4.3: Statistical summary of the variables used for evaluating the speeches impact of the S&P 500 observations on the day before with respect those in which the talks are delivered.

respect to the previous case. This is due to the higher concentration of observations in the left side of the distributions.

The visual inspection of Figures 4.5 and 4.6 leads to similar conclusions.

	Economic Rel. Freq.	Economic Abs. Freq. Norm.	Prices Norm.	Volume Norm.	Ret. Norm.
N. Obs.	379	379	379	379	379
Max	0.075	0.862	1.000	0.708	0.814
Min	0.003	0.001	0.000	0.000	0.517
Median m	0.018	0.062	0.048	0.005	0.678
Mean μ	0.020	0.099	0.188	0.081	0.677
RMS	0.023	0.142	0.294	0.169	0.678
St. Dev. σ	0.011	0.102	0.227	0.149	0.033
Var.	0.000	0.010	0.051	0.022	0.001
Sd. Err.	0.001	0.005	0.012	0.008	0.002
Skewness	1.605	2.397	1.297	1.988	-0.259
Kurtosis	6.754	12.696	3.637	6.061	8.375
μ/σ	1.847	0.965	0.826	0.545	20.752
$3(\mu - m)/\sigma$	0.673	1.061	1.842	1.532	-0.135

Table 4.4: Statistical summary of the variables used for evaluating the speeches impact of the S&P 500 observations on the day ahead with respect those in which the talks are delivered

	One day back of S&P 500 observation		Contemporaneous date		One day ahead of S&P 500 observation	
	τ	p-value	τ	p-value	τ	p-value
Norm. Ret. and Rel. Freq.	-0.014	0.347	-0.013	0.362	0.01	0.387
Norm. Ret. and Abs. Freq. Norm.	-0.006	0.436	0.007	0.43	0.022	0.263
Norm. Closing Price and Rel. Freq.	0.089	0.005	0.087	0.009	0.09	0.004
Norm. Closing Price and Abs. Freq. Norm.	0.106	0.001	0.104	0.003	0.106	0.001
Norm. Vol. and Rel. Freq.	0.112	0.001	0.105	0.002	0.107	0.001
Norm. Vol. and Abs. Freq. Norm.	0.12	0	0.116	0.001	0.114	0

Table 4.5: The Kendall's τ estimations. The columns are divided in groups on the basis of the selection timing of S&P 500 observation. The first couple is referred to the case of S&P 500 variable on the same dates of Presidents' speeches, while the second and the third ones are the case of the subsequent and the previous day of Presidents' speeches, respectively.

Histograms present asymmetries with right tails which are fatter than the ones in the previous cases. The kurtosis are remarkably large for the cases of normalized volumes, where the tallest bins of the graphs contain the majority of the values that fall around zero. Such an occurrence affects all the statistical position indicators and lowers the variances, which is smaller than that of the normalized prices. The presence of outliers is relevant also in this case.

Figure 4.7 shows very different behaviors with respect to the other considered variables. This outcome supports the evidence that the normalized returns show symmetric distributions, with skewness values very close to zero and means and medians almost coinciding. The variances are tiny inasmuch the distributions are concentrated around the centers with thin tails. From a visual inspection, it is possible to note that the outliers are present in all the histograms, but the cases of the day before Presidents' talks dates are associated to slightly fatter right sides.

Table 4.5 shows that the τ estimations are not statically significant when the S&P 500 normalized returns are involved into the evaluations, as it is in couples (a) and (b).

All the other cases have outstanding statistical significance levels below 1%. The cases (c) and (d) present a positive influence of the economic locutions presence on the normalized closing prices. Specifically, the smallest positive rank correlations appears in the contemporaneous cases, while the effects on the day before and the day after the speeches delivery dates are quite similar and slightly bigger than the contemporaneous case. For some cases in (e) and (f) the p-values are closer to zero than the ones of the previous couples. So, τ estimations are extraordinarily significant and they are very close for the first two instances (contemporaneous and one day ahead with respect to the speeches dates). In (e), the positive correlations are 10.5% for the contemporaneous effects and 10.7% for the posterior effects. A noticeable 11.2% is registered by investigating the relationship between the volumes of the day before the speeches. In (f) the situation in almost the same, but the magnitude of the positive correlations reaches 12% when the analysis is performed on the day before the speeches dates. This suggests an interesting predictive capacity.

The rank-rank positive correlations of the economic locutions presence into the speeches and the S&P 500 index go from a minimum of 8.7% for the contemporaneous case of (c) to a maximum of 12% of the (f) result when the S&P 500 volumes observations are registered the days before the speeches delivery. Such estimated correlations, along with the respective large statistical significance, seems to confirm the intuitive fact that the extreme power of the US Presidents inevitably affects the markets even by means of a talk with an economic content. In this light, the positive correlations of the volumes with the presence of economic terms –cases (e) and (f) – can be considered as a measure of the attention paid by the traders to the US Presidents' economics messages. From this overview, we can partially explain the changes in prices occurred into the considered time windows (from one day before the speeches to one day after) due to the informative capacity of the Presidents' talks. Indeed the correlations change along the time with a decrease in correspondence of the days of the speeches ('at the date' comparison). In those days the changes in prices are less influenced by the presidents' speeches because of the uncertainty of the message that is usually delivered when the market is closed, while the days before the speeches, the prices move in accord to the expectations on the talks' content (in case the speeches were planned for that specific days in the political agenda).While, in the days after the prices' adjustments occur for correcting the traders' expectations on the messages or for incorporate the information into the prices. The correlation measured on the volumes further confirm it, especially for the case (e). In the days before that the speeches are delivered there are noticeable increment of positive correlation between volumes and economic terms relative frequencies.

Let us now consider the distances reported in Table 4.6. The distances calculated with eqs. (4.3) and (4.4) are useful to identify the extremes of the range of variation of the deviation series. Indeed, by looking at the outcomes of eq. (4.3), one can note that the distance estimations are quite big and homogeneous, whilst equation (4.4) shows big differences within the estimations and also null values.

Distances in eqs. (4.5) and (4.6) underline two different aspects. The former penalizes the extreme differences points and the second amplifies them when they are greater than one through the square in the formula. So we respectively have a prudent measure that underestimates the differences and another that overestimates them one which are greater than one.

Cases (d) and (f) do not show deviations across the columns of Table 4.6 when eqs. (4.3), (4.4) and (4.5) are used. This happens also for cases (c) and (e) when eq. (4.4) is employed and for (a) in case of formula (4.5).

The outcomes of (b), (c) and (e) in the case of distance in (4.5) have tiny changes across the columns while the other couples are equal. This leads to the conclusion that the differences between variables selected in different dates do not change the common behaviors that much. Furthermore, couples (e) and (f) have minimum distances when formula (4.5) is used, proving a similar behavior over time. Such outcomes validate the findings of the Kendall's τ for the same pairs of variables (see Table 4.5). It reinforces the idea that the S&P 500 volumes are affected by the presence of economic terms into the speeches.

Distances taken with formula (4.6) shows interesting results for the couples (a), (b), (c), (d) and (f). They have the minimum Euclidean distances in the contemporaneous dates of observations ('at the date' case), the maximum when the S&P 500 data is considered on the day after the speeches and a small inflection if the observations are registered the day before. It can be explained with the same logic used for the Kendall's τ correlation. Namely, the traders change their positions the days before that the speeches have been stated affecting prices and even more the volumes. The minimum Euclidean distances confirm that the S&P 500 volumes are the series closest to the frequencies series of the economic terms, validating the conclusion derived from the results of the τ coefficients estimations. On the other hand, by using both formulas (4.5) and (4.6), one has that the most distant series from the economic locutions frequencies (normalized absolute and relative) are the normalized daily returns series. This finding is coherent with the scarce statistical significance obtained when the returns are involved in the τ estimations. This conclusion is reinforced by the noticeable cases of distance as with eq. (4.4), where the unique couples that are not zero are the ones where the normalized returns are involved.

Table 4.7 reports the computations of the Shannon's entropy through formula (4.7). The highest entropy is registered with the series of the economic terms' relative frequencies, while the lowest is obtained in the case of the S&P 500's volumes.

The series of normalized absolute and relative frequencies exhibit entropy values that are the closest to the entropy of the returns. This suggests that the disorder of such series are similar, hence leading to similar shapes of the empirical distributions of frequencies and returns, in all the instances of contemporaneous, one day ahead and one day before. If we compare such an outcome with the previous results of the Kendall's τ and distances, one can argue that frequencies can be viewed as a proxy of returns in terms of distribution and main statistical properties but the behaviors of frequencies and returns are not simultaneous. Substantially, from the entropies it is not possible to state whether returns and frequencies deviate or not from common paths. This is especially true considering the results from rank-rank correlations and measurement of distances. The likelihood between entropies is more evident in the case of normalized absolute frequencies. The most remarkable deviations can be observed between frequencies and volumes. This finding can be read in the light of what we had for Kendall and distances. Specifically, volumes seem to show a positive correlation with frequencies when rank-rank analysis is performed, but the overall distributions of volumes and frequencies exhibit different shapes. Thus, one can use the frequencies as a proxy of the volumes in terms of evolution but not in terms of statistical features of the related empirical distribution.

Measure used	Data used	One day back of S&P500 obs.	Contemporaneous dates	One day ahead of S&P500 obs.
Eq. (4.3)	Abs. Freq. Norm. and Norm. Closing Price	0.91	0.91	0.91
Eq. (4.3)	Abs. Freq. Norm. and Norm. Vol.	0.81	0.81	0.81
Eq. (4.3)	Abs. Freq. Norm. and Norm. Ret.	0.75	0.77	0.77
Eq. (4.3)	Rel. Freq. and Norm. Closing Price	0.96	0.95	0.97
Eq. (4.3)	Rel. Freq. and Norm. Vol.	0.59	0.65	0.68
Eq. (4.3)	Rel. Freq. and Norm. Ret.	0.78	0.78	0.78
Eq. (4.4)	Abs. Freq. Norm. and Norm. Closing Price	0	0	0
Eq. (4.4)	Abs. Freq. Norm. and Norm. Vol.	0	0	0
Eq. (4.4)	Abs. Freq. Norm. and Norm. Ret.	0.14	0.13	0.08
Eq. (4.4)	Rel. Freq. and Norm. Closing Price	0	0	0
Eq. (4.4)	Rel. Freq. and Norm. Vol.	0	0	0
Eq. (4.4)	Rel. Freq. and Norm. Ret.	0.43	0.49	0.46
Eq. (4.5)	Abs. Freq. Norm. and Norm. Closing Price	0.17	0.17	0.17
Eq. (4.5)	Abs. Freq. Norm. and Norm. Vol.	0.12	0.12	0.12
Eq. (4.5)	Abs. Freq. Norm. and Norm. Ret.	0.58	0.57	0.58
Eq. (4.5)	Rel. Freq. and Norm. Closing Price	0.16	0.17	0.16
Eq. (4.5)	Rel. Freq. and Norm. Vol.	0.08	0.09	0.09
Eq. (4.5)	Rel. Freq. and Norm. Ret.	0.65	0.65	0.65
Eq. (4.6)	Abs. Freq. Norm. and Norm. Closing Price	4.91	4.55	4.92
Eq. (4.6)	Abs. Freq. Norm. and Norm. Vol.	3.2	3.17	3.34
Eq. (4.6)	Abs. Freq. Norm. and Norm. Ret.	11.36	10.59	11.38
Eq. (4.6)	Rel. Freq. and Norm. Closing Price	5.36	5.01	5.37
Eq. (4.6)	Rel. Freq. and Norm. Vol.	2.86	2.89	3.04
Eq. (4.6)	Rel. Freq. and Norm. Ret.	12.58	11.82	12.61

Table 4.6: The different measures of distance described in Section 4.3 are computed and reported. The last three columns are referred to different selections of the S&P 500 variables. The third last is about the case of S&P 500 taken on the same dates in which the Presidents' speeches have been delivered, while the second to last and the last are the cases of the subsequent and the previous day with respect to the dates in which the Presidents have spoken, respectively.

	One day before the speeches' dates	Contemporaneous dates	One day after the speeches' dates
Terms' rel. freq.	6.76	6.79	6.78
Terms' abs. freq. norm.	6.15	6.19	6.15
S&P500 Closing price norm.	5.88	5.81	5.83
S&P500 Vol. norm.	4.26	4.27	4.24
S&P500 Ret. norm.	6.2	6.07	6.32

Table 4.7: In this table is reported Shannon's entropy for each series is reported. The columns distinguish the different data selection made in accord to the selected dates. The first column points to the S&P 500 observations registered the same days in which the Presidents have stated their talks, while the second and the third one are about the series of S&P 500 observations registered the day after and the day before the Presidents' talks.

4.6 Conclusive remarks

Some conclusions can be derived by looking at the results of the performed analysis.

The economic content in the talks of the US Presidents have an impact on the Standard and Poor's 500 index and the impact magnitude depends on the S&P 500's variables. In accord to the findings here presented, the S&P 500's volumes are the most sensible to the speeches with presence of locutions belonging to the economics glossary. Specifically, the τ coefficient indicates a positive and significant correlation and the measures of the distance present a good degree of synchronicity between the series. Furthermore the Shannon entropy indicates similarities between the behavior of the volumes and the frequencies of the economic terms, giving another confirmation.

The S&P 500's closing prices have lower rank–rank correlation with the economic locutions frequencies (for both relative and absolute) than the volumes while there is evidence of wide distances between these variables. The closing prices' entropies are closer to those related to the frequencies series of economic terms. Therefore the conclusions on the relationship of closing prices and the economic content of the Presidents' speeches are similar to them reached for the volumes. Indeed the S&P 500 prices changes before and after the days in which the speeches are delivered, following the steps of the volumes.

Lastly the S&P 500's daily returns do not show significant changes related to the economic locutions presence into the speeches, especially when the Kendall's τ is employed. The different measurements of the distance prove that the returns are always far from the series of the glossary terms' frequencies, and this could partially justify the absence of statistical significance in the τ estimations. The outcomes are different when the Shannon's entropy is evaluated. Indeed, the entropies of the returns series are close to the frequencies terms series. This suggests that, even if an instantaneous relationship is not manifested, there could be a relationship under the empirical distribution point of view.

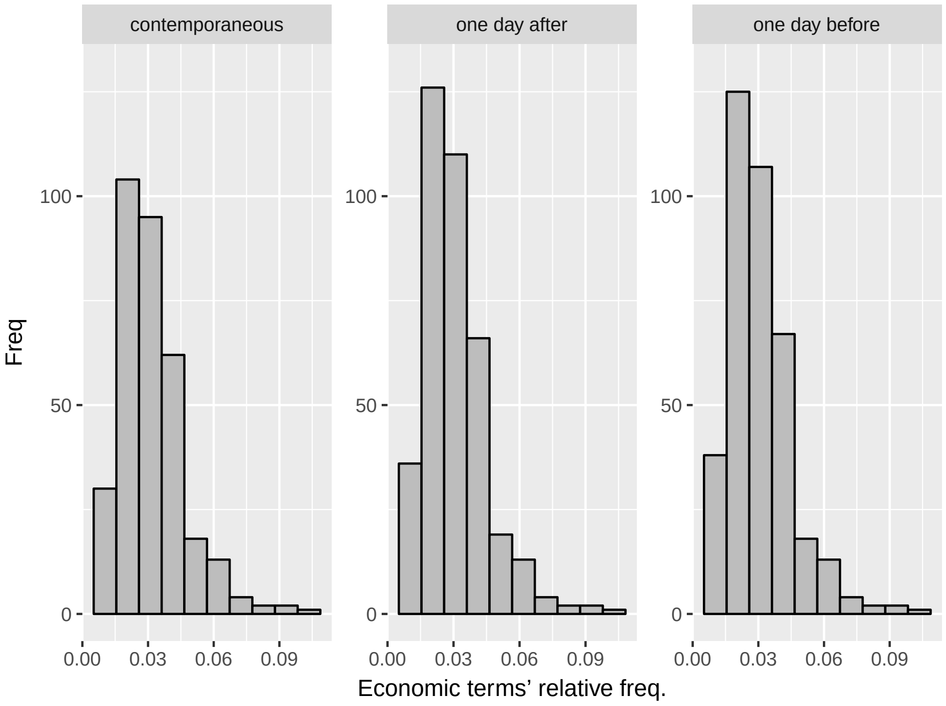

Figure 4.3: Histograms of the relative frequencies of the economic terms. The differences between sub-figures are given by the dates selection, namely the speeches that do not have corresponding S&P 500 observations are excluded. For example, when a speech is stated on Friday, the day after it does not have a correspondent S&P 500 observation because the market is closed. Consequently that particular talk is cancelled.

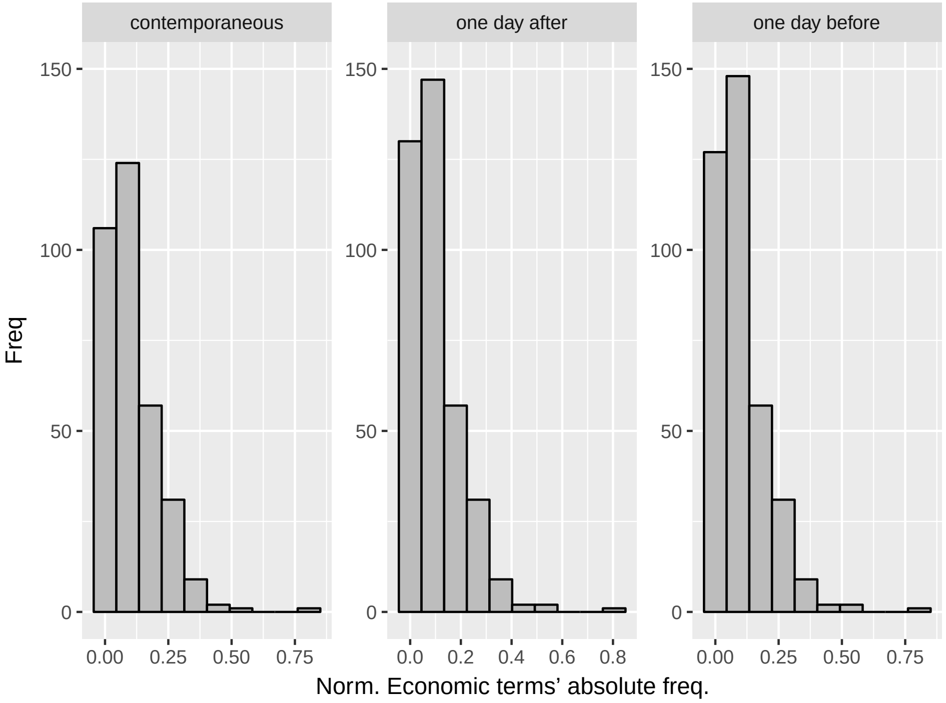

Figure 4.4: Histograms of the normalized absolute frequencies of the economics terms. The differences between sub-figures are given by the dates selection, namely the speeches that do not have corresponding S&P 500 observations are excluded. For example, when a speech is stated on Friday, the day after it does not have a correspondent S&P 500 observation because the market is closed. Consequently that particular talk is cancelled.

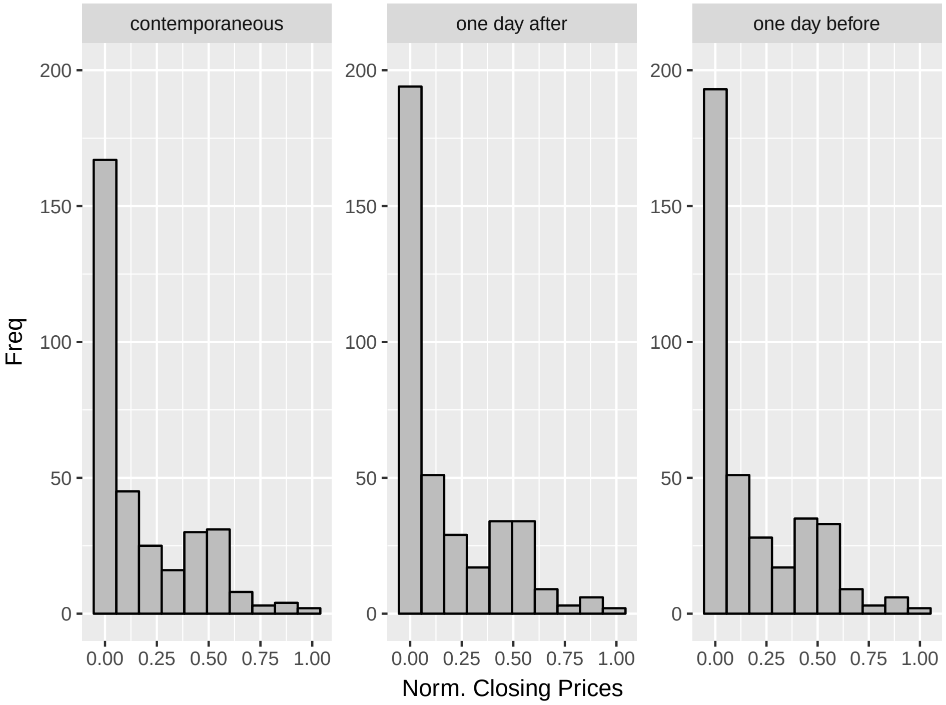

Figure 4.5: Histograms of the S&P 500's normalized daily closing prices

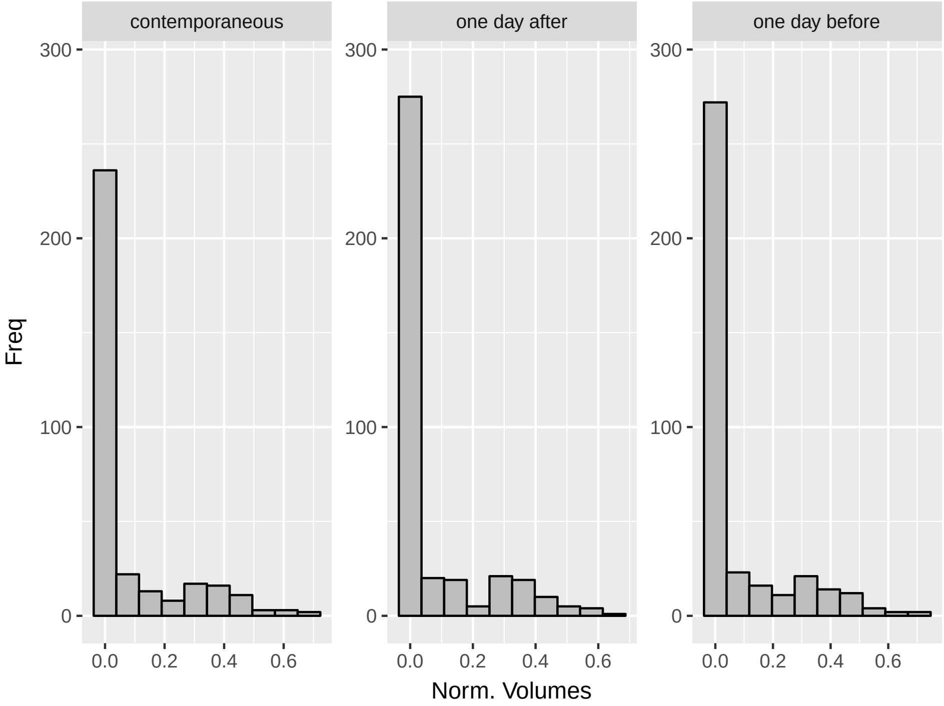

Figure 4.6: Histograms of the S&P 500's normalized daily volumes

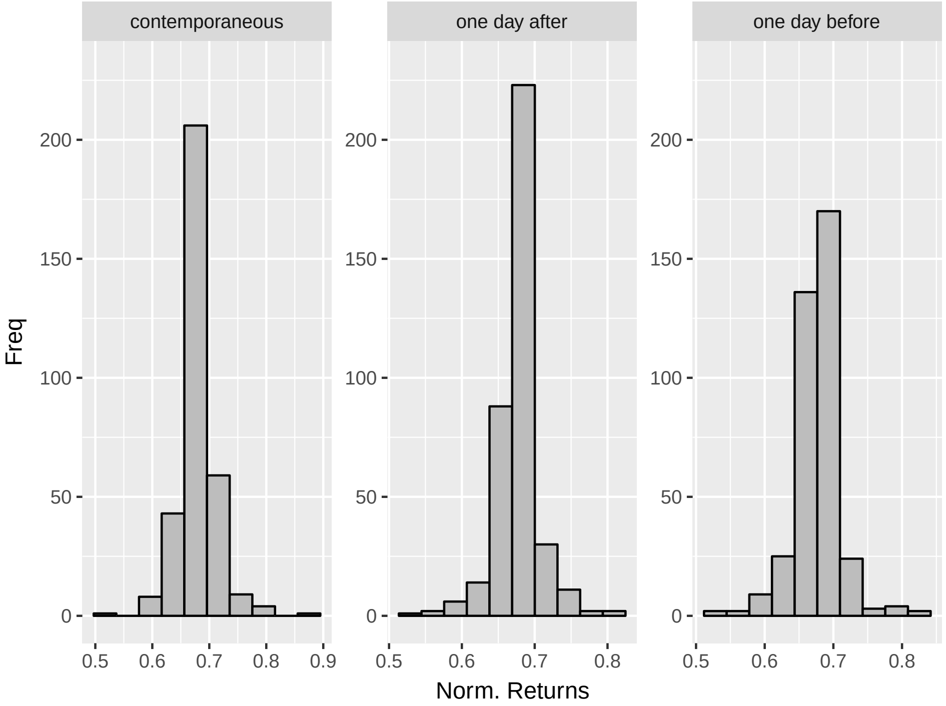

Figure 4.7: Histograms of the S&P 500's normalized daily returns

Bibliography

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[2] B. Aguilar-San Juan and L. Guzman-Vargas. Earthquake magnitude time series: scaling behavior of visibility networks. *The European Physical Journal B*, 86(11):454, 2013.

[3] C. Alduy. *Ce qu'ils disent vraiment. Les politiques pris aux mots.* Le Seuil, 2017.

[4] E. G. Altmann, G. Cristadoro, and M. Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

[5] A. Amato and F. Mele. Performance of the ingv national seismic network from 1997 to 2007. *Annals of Geophysics*, 51(2-3):417–431, 2008.

[6] American-Israeli Cooperative Enterprise. Jewish virtual library, 1993.

[7] W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

[8] M. Ausloos. Equilibrium and dynamic methods when comparing an English text and its Esperanto translation. *Physica A: Statistical Mechanics and its Applications*, 387(25):6411–6420, 2008.

[9] M. Ausloos. Punctuation effects in English and Esperanto texts. *Physica A: Statistical Mechanics and its Applications*, 389(14):2835–2840, 2010.

[10] M. Ausloos. Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. *Physical Review E*, 86(3):031108, 2012.

[11] M. Ausloos. Measuring complexity with multifractals in texts. Translation effects. *Chaos, Solitons & Fractals*, 45(11):1349–1357, 2012.

[12] M. Ausloos. A scientometrics law about co-authors and their ranking: the co-author core. *Scientometrics*, 95(3):895–909, 2013.

[13] M. Ausloos and R. Cerqueti. A universal rank-size law. *PloS one*, 11(11):e0166011, 2016.

[14] M. Ausloos, R. Cloots, A. Gadomski, and N. K. Vitanov. Ranking structures and rank–rank correlations of countries: The fifa and uefa cases. *International Journal of Modern Physics C*, 25(11):1450060, 2014.

[15] M. Ausloos, O. Nedic, A. Fronczak, and P. Fronczak. Quantifying the quality of peer reviewers through Zipfs law. *Scientometrics*, 106(1):347–368, 2016.

[16] R. L. Axtell. Zipf distribution of US firm sizes. *Science*, 293(5536):1818–1820, 2001.

[17] S. R. Baker, N. Bloom, and S. J. Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

[18] A. Baron, P. Rayson, and D. Archer. Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67, 2009.

[19] S. R. Bentes, R. Menezes, and D. A. Mendes. Long memory and volatility clustering: Is the empirical evidence consistent across stock markets? *Physica A: Statistical Mechanics and its Applications*, 387(15):3826–3830, 2008.

[20] C. Bentz, D. Kiela, F. Hill, and P. Buttery. Zipf's law and the grammar of languages: A quantitative study of old and modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10(2):175–211, 2014.

[21] M. Bishop. *Essential economics: an A to Z guide*, volume 22. John Wiley & Sons, 2009.

[22] B. Blasius and R. Tönjes. Zipfs law in the popularity distribution of chess openings. *Physical Review Letters*, 103(21):218701, 2009.

[23] G. Bottazzi, D. Pirino, and F. Tamagni. Zipf law and the firm size distribution: a critical discussion of popular estimators. *Journal of Evolutionary Economics*, 25(3):585–610, 2015.

[24] A. S. Calude and M. Pagel. How do we use language? shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1567):1101–1107, 2011.

[25] A. Carretta, V. Farina, D. Martelli, F. Fiordelisi, and P. Schwizer. The impact of corporate governance press news on stock market returns. *European financial management*, 17(1):100–119, 2011.

[26] R. Cerqueti and M. Ausloos. Cross ranking of cities and regions: population versus income. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(7):P07002, 2015.

[27] R. Cerqueti and M. Ausloos. Evidence of economic regularities and disparities of Italian regions from aggregated tax income size data. *Physica A: Statistical Mechanics and its Applications*, 421:187–207, 2015.

[28] L. Chiaraluce, R. Di Stefano, E. Tinti, L. Scognamiglio, M. Michele, E. Casarotti, M. Cattaneo, P. De Gori, C. Chiarabba, G. Monachesi, et al. The 2016 central italy seismic sequence: A first look at the mainshocks, aftershocks, and source models. *Seismological Research Letters*, 88(3):757–771, 2017.

[29] L. Débowski. Zipf's law against the text size: a half-rational model. *Glottometrics*, 4:49–60, 2002.

[30] Z. Dimitrova and M. Ausloos. Primacy analysis of the system of Bulgarian cities. *Open Physics*, 13(1):218–225, 2015.

[31] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison

of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[32] P. Dorčák, F. Pollák, and S. Szabo. Analysis of the possibilities of improving an online reputation of public institutions. *IDIMT-2014, Sept. 10–12. Poděbrady: IDIMT Networking Societies-Cooperation and Conflict 22ndInterdisciplinary Information and Management Talks*, pages 275–281, 2014.

[33] A. A. Dragulescu. *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files*, 2014. R package version 0.5.7.

[34] A. A. Dragulescu. *xlsxjars: Package required POI jars for the xlsx package*, 2014. R package version 0.6.1.

[35] S. Drożdż, P. Oświcimka, A. Kulig, J. Kwapień, K. Bazarnik, I. Grabska-Gradzińska, J. Rybicki, and M. Stanuszek. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32–44, 2016.

[36] R. A. Fairthorne. Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation*, 61(2):171–193, 2005.

[37] V. Farina, A. Parisi, and U. Pomante. Economics blogs sentiment and asset prices. *International Journal of Finance & Economics*, 22(4):341–351, 2017.

[38] R. Ferrer-i Cancho and B. Elvevåg. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One*, 5(3):e9411, 2010.

[39] M. Fujita and J.-F. Thisse. The formation of economic agglomerations: old problems and new perspectives. *Economics of cities: Theoretical perspectives*, pages 3–73, 2000.

[40] G. P. C. Fung, J. X. Yu, and W. Lam. News sensitive stock trend prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 481–493. Springer, 2002.

[41] X. Gabaix and Y. M. Ioannides. The evolution of city size distributions. *Handbook of regional and urban economics*, 4:2341–2378, 2004.

[42] D. Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[43] J. D. Gibbons. *Nonparametric measures of association*. Number 91. Sage, 1993.

[44] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[45] T. S. Hare. Structural relationships within and among aztec communities and polities. *The archaeology of communities: A New World perspective*, pages 78–101, 2000.

[46] J. Hausser and K. Strimmer. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*, 2014. R package version 1.2.1.

[47] T. Hendershott, D. Livdan, and N. Schürhoff. Are institutions informed about news? *Journal of Financial Economics*, 117(2):249–287, 2015.

[48] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

[49] S.-M. Huang, D. C. Yen, L.-W. Yang, and J.-S. Hua. An investigation of Zipf's law for fraud detection (dss# 06-10-1826r (2)). *Decision Support Systems*, 46(1):70–83, 2008.

[50] R. F. i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

[51] INGV. Definitions, 2018.

[52] Y. Ioannides and S. Skouras. Us city size distribution: Robustly pareto, but only in the tail. *Journal of Urban Economics*, 73(1):18–29, 2013.

[53] Y. M. Ioannides and H. G. Overman. Zipfs law for cities: an empirical examination. *Regional science and urban economics*, 33(2):127–137, 2003.

[54] S. C. Jaumé. Changes in earthquake size-frequency distributions underlying accelerating seismic moment/energy release. *Geocomplexity and the Physics of Earthquakes*, pages 199–210, 2000.

[55] M. Jefferson. The law of the primate city. *Geographical Review*, 29(2):226–232, 1939.

[56] M. L. Jockers. *Text analysis with R for students of literature*. Springer, 2014.

[57] J. L. Juergens and L. Lindsey. Getting out early: An analysis of market making activity at the recommending analyst's firm. *The Journal of Finance*, 64(5):2327–2359, 2009.

[58] Y. Y. Kagan. Earthquake size distribution: Power-law with exponent $\beta \equiv$ 12? *Tectonophysics*, 490(1-2):103–114, 2010.

[59] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[60] W. H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[61] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy:fat tails with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[62] C. Largeron, C. Moulin, and M. Géry. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 924–928. ACM, 2011.

[63] D. Lavalette. Facteur dimpact: impartialité ou impuissance. *Report, INSERM U*, 350:91405, 1996.

[64] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Mining of concurrent text and time series. In *KDD-2000 Workshop on Text Mining*, volume 2000, pages 37–44, 2000.

[65] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

[66] M. Levene, J. Borges, and G. Loizou. Zipf's law for web surfers. *Knowledge and Information Systems*, 3(1):120–129, 2001.

[67] W. Li and Y. Yang. Zipf's law in importance of genes for cancer classification using microarray data. *Journal of Theoretical Biology*, 219(4):539–551, 2002.

[68] E. T. Lim. Five trends in presidential rhetoric: An analysis of rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, 32(2):328–348, 2002.

[69] M. I. Lourakis. A brief description of the Levenberg-Marquardt algorithm implemented by levmar. *Foundation of Research and Technology*, 4(1), 2005.

[70] T. Maillart, D. Sornette, S. Spaeth, and G. Von Krogh. Empirical tests of Zipfs law mechanism in open source Linux distribution. *Physical Review Letters*, 101(21):218701, 2008.

[71] B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, and R. B. Davis. Zipf's law, music classification, and aesthetics. *Computer Music Journal*, 29(1):55–69, 2005.

[72] B. Mandelbrot. An informational theory of the statistical structure of language. *Communication theory*, 84:486–502, 1953.

[73] B. Mandelbrot. On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 12:190–219, 1961.

[74] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[75] A. Marchetti, M. G. Ciaccio, A. Nardi, A. Bono, F. M. Mele, L. Margheriti, A. Rossi, P. Battelli, C. Melorio, B. Castello, et al. The italian seismic bulletin: strategies, revised pickings and locations of the central italy seismic sequence. *Annals of Geophysics*, 59, 2016.

[76] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[77] V. J. Matlaba, M. J. Holmes, P. McCann, and J. Poot. A century of the evolution of the urban system in brazil. *Review of Urban & Regional Development Studies*, 25(3):129–151, 2013.

[78] A. M. Mayda and G. Peri. The economic impact of us immigration policies in the age of trump. *Economics and Policy in the Age of Trump*, 69, 2017.

[79] A. McLeod. *Kendall: Kendall rank correlation and Mann-Kendall trend test*, 2011. R package version 2.2.

[80] B. McNair. *An introduction to political communication*. Taylor & Francis, 2017.

[81] M. S. Mega, P. Allegrini, P. Grigolini, V. Latora, L. Palatella, A. Rapisarda, and S. Vinciguerra. Power-law time distribution of large earthquakes. *Physical Review Letters*, 90(18):188501, 2003.

[82] M. Melucci. On rank correlation in information retrieval evaluation. In *ACM SIGIR Forum*, volume 41, pages 18–33. ACM, 2007.

[83] D. Milizia. In, out, or half way? The European attitude in the speeches of British leaders. *Lingue e Linguaggi*, 11:157–175, 2014.

[84] Miller Center. Inaugural address, 1925.

[85] Miller Center. Sixth annual message, 1928.

[86] Miller Center. Campaign speech in indianapolis, indiana., 1932.

[87] Miller Center. Press conference in the east room, 1966.

[88] Miller Center. First annual message, 1981.

[89] Miller Center. Remarks honoring the vietnam wars unknown soldier, 1984.

[90] Miller Center. Speech on foreign policy, 1988.

[91] Miller Center. Remarks at the democratic national convention, 1996.

[92] Miller Center. Remarks at the democratic national convention, 2004.

[93] Miller Center. Address to the united nations, 2010.

[94] Miller Center. Remarks in eulogy for the honorable reverend clementa pickney, 2015.

[95] Miller Center. 2016 state of the union address, 2016.

[96] R. Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.

[97] M.-A. Mittermayer and G. F. Knolmayer. Newscats: A news categorization and trading system. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1002–1007. Ieee, 2006.

[98] M. A. Montemurro. Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.

[99] M. Moretti, B. Baptie, and M. Segou. Sismiko: emergency network deployment and data sharing for the 2016 central italy seismic sequence. *Annals of Geophysics*, 59(5), 2017.

[100] L. A. Mullen, K. Benoit, O. Keyes, D. Selivanov, and J. Arnold. Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, 3:655, 2018.

[101] S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis. *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons, 2014.

[102] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670, 2014.

[103] G. Natale, F. Musmeci, and A. Zollo. A linear intensity model to investigate the causal relation between calabrian and north-aegean earthquake sequences. *Geophysical Journal International*, 95(2):285–293, 1988.

[104] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[105] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[106] A. Nikfarjam, E. Emadzadeh, and S. Muthaiyah. Text mining approaches for stock market prediction. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 4, pages 256–260. IEEE, 2010.

[107] A. Pak and P. Paroubek. Twitter for sentiment analysis: When language resources are not available. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 111–115. IEEE, 2011.

[108] C. Papadimitriou, K. Karamanos, F. Diakonos, V. Constantoudis, and H. Papageorgiou. Entropy analysis of natural language written texts. *Physica A: Statistical Mechanics and its Applications*, 389(16):3260–3266, 2010.

[109] G. Peng. Zipfs law for chinese cities: rolling sample regressions. *Physica A: Statistical Mechanics and its Applications*, 389(18):3804–3813, 2010.

[110] S. T. Piantadosi. Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014.

[111] C. M. Pinto, A. M. Lopes, and J. T. Machado. A review of power laws in real life phenomena. *Communications in Nonlinear Science and Numerical Simulation*, 17(9):3558–3578, 2012.

[112] G. Polizzi. Giacomo leopardi. in: Il contributo italiano alla storia del pensiero, ottava appendice. *Istituto della Enciclopedia Italiana Fondata da Giovanni Treccani, Roma*, 1(8):426–433, 2012.

[113] I.-I. Popescu, G. Altmann, and R. Köhler. Zipfs lawanother view. *Quality & Quantity*, 44(4):713–731, 2010.

[114] T. Preis, H. S. Moat, and H. E. Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3:srep01684, 2013.

[115] L. Romashkova and A. Peresan. Analysis of italian earthquake catalogs in the context of intermediate-term prediction problem. *Acta Geophysica*, 61(3):583–610, 2013.

[116] K. T. Rosen and M. Resnick. The size distribution of cities: an examination of the pareto law and primacy. *Journal of Urban Economics*, 8(2):165–186, 1980.

[117] O. A. Rosso, H. Craig, and P. Moscato. Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*, 388(6):916–926, 2009.

[118] A. Rovenchak and S. Buk. Part-of-speech sequences in literary text: Evidence from ukrainian. *Journal of Quantitative Linguistics*, 25(1):1–21, 2018.

[119] A. Saichev and D. Sornette. Power law distribution of seismic rates: theory and data analysis. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(3):377–401, 2006.

[120] D. Schorlemmer, F. Mele, and W. Marzocchi. A completeness analysis of the national seismic network of italy. *Journal of Geophysical Research: Solid Earth*, 115(B4), 2010.

[121] R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[122] M. Scott. Wordsmith tools version 4: computer program, 2004.

[123] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[124] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[125] Y. Shynkevich, T. M. McGinnity, S. Coleman, and A. Belatreche. Predicting stock price movements based on different categories of news articles. In *Computational Intelligence, 2015 IEEE Symposium Series on*, pages 703–710. IEEE, 2015.

[126] J. Stallings, E. Vance, J. Yang, M. W. Vannier, J. Liang, L. Pang, L. Dai, I. Ye, and G. Wang. Determining scientific impact using a collaboration index. *Proceedings of the National Academy of Sciences*, 110(24):9680–9685, 2013.

[127] S. Stemler. An overview of content analysis. *Practical assessment, research & evaluation*, 7(17):137–146, 2001.

[128] C. Suteanu, L. Liucci, and L. Melelli. The central italy seismic sequence (2016): Spatial patterns and dynamic fingerprints. *Pure and Applied Geophysics*, pages 1–24, 2018.

[129] R. Sutter. The united states and asia in 2017: The impact of the trump administration. *Asian Survey*, 58(1):10–20, 2018.

[130] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *ACM SIGIR Forum*, volume 51, pages 10–17. ACM, 2018.

[131] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[132] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[133] S. M. Toorawa. Hapless hapaxes and luckless rhymes: The qur'an as literature. *Religion & Literature*, 41(2):221–227, 2009.

[134] L. Vavreck. *The message matters: The economy and presidential campaigns*. Princeton University Press, 2009.

[135] S. Vejdemo and T. Hörberg. Semantic factors predict the rate of lexical replacement of content words. *PloS one*, 11(1):e0147924, 2016.

[136] A. J. Venables. The spatial economy: Cities, regions, and international trade. 1999.

[137] C.-J. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang. Financial sentiment analysis for risk prediction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 802–808, 2013.

[138] H. Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2016. R package version 0.3.2.

[139] H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2018. R package version 1.3.0.

[140] H. Wickham, J. Hester, and J. Ooms. *xml2: Parse XML*, 2018. R package version 1.2.0.

[141] Z. Wu. Frequency–size distribution of global seismicity seen from broad-band radiated energy. *Geophysical Journal International*, 142(1):59–66, 2000.

[142] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang. Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 3, pages 2720–2725. IEEE, 1998.

[143] F. Yoshi. Zipf's law in firms bankruptcy. *Physica A: Statistical Mechanics and its Applications*, 2004.

[144] D. H. Zanette. Zipf's law and the creation of musical context. *Musicae Scientiae*, 10(1):3–18, 2006.

[145] Y.-B. Zhou, L. Lü, and M. Li. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics*, 14(3):033033, 2012.

[146] G. K. Zipf. The psycho-biology of language. 1935.

[147] G. K. Zipf. Human behaviour and the principle of least-effort. Cambridge MA edn, 1949.