UNIVERSITY OF TRENTO - Italy

Doctoral School of Social Sciences
Doctoral programme in
Economics and Management

Four Essays on Machine Learning in Microeconomics and Macroeconomics

a dissertation submitted in partial fulfillment of the requirements for the doctoral degree (Ph.D.) in Economics and Management

Giacomo Caterini
January 2020

Supervisor: Professor Edoardo Gaffeo (University of Trento)

Doctoral Committee
Prof. Matteo Ploner (University of Trento)
Prof. Sandra Paterlini (University of Trento)
Prof. Roberto Tamborini (University of Trento)

Acknowledgements

To begin, I want to thank my Supervisor, Prof. Edoardo Gaffeo, for several reasons. First of all, for his role as guide at the beginning of my career, leading me toward the choice of my way. He encouraged me to discover the realm of machine learning and he probably understood -before I did- that through Text Mining I had the chance to apply my passion for the literature to the Economic discipline, "playing" with words. Secondly, I want to thank Edoardo for his efforts aimed at fulfilling bureaucratic requirements needed to complete my Ph.D. program in time. For the same reason, my thanks go to the Doctoral School of Social Sciences as a whole. Finally, and most important, I thank Edoardo for his support during hard moments of my career and personal life.

I want to thank the referees, Prof. František Dařena and Dr. Juri Marcucci, for their suggestions that will surely be extremely useful when I will submit my papers to scientific journals.

I thank Dr. Georg Lun, the director of the Institute for Economic Research (IER) of the Chamber of Commerce of Bolzano. The Institute financed my Ph.D. scholarship and provided the data that I used in Chapter 3 and 4 of this work. I specially want to thank Luciano Partacini and Mattias Martini, who provided their precious support during the time that I spent at IER. I also want to thank Silvia Berlanda, Cristina Bagante, Nicola Riz, Alessio Tomelleri and all the people I met at IER.

I want to express my gratitude to professors that I met at the University of Trento, including Luciano Andreozzi, Emanuele Taufer, Flavio Bazzana, Luca Erzegovesi, Luigi Mittone, Matteo Ploner and Carlo Fezzi. I specially thank Prof. Roberto Tamborini, who has been my macroeconomic mentor together with Prof. Gaffeo. Moreover, I'm grateful to Enrico Zaninotto, who allowed me to attend several editions of the Trento Summer School in AED, which revealed to be fundamental experiences for me as an economist.

Over time, several researchers provided useful comments: among them, I want to mention Enrico Zaninotto, Cathy Yi-Hsuan Chen, Agostino Di Ciaccio, Giuseppe Riccardi, Stefano Varotti and the participants to the 29th Conference on Big Data Econometrics. I also want to thank Prof. Proietti and Dr. Cristofaro for the comments that they gave during my visiting at the University of Tor Vergata.

I want to express my gratitude to all the colleagues and friends which have stood by me in the last four years, in particular: Amir Maghssudipour, Giulio Galdi, Alessia Dorigoni, Filippo Santi, Luca Bortolotti, Francesca Nuzzo, Enrico Cristofoletti, Lucia Pederiva and Lucio Gobbi.

If I have been able to complete my Ph.D., to overcome difficulties, to keep calm and to preserve my mental health, the credit for this is due to Cristina, who supported me when I faced the hardest period of my life. I will be grateful to her forever.

Finally, my gratitude goes to my parents and grandparents, which allowed me to delve my education and which always granted the freedom to err.

Introduction

In the last decade, the role of Machine Learning in Economics is becoming preeminent for a wide range of applications including accounting, marketing, finance, official statistics and economic forecasting. This is mainly due to the emergence of "Big Data": such a phenomenon refers to the explosion in the quantity and quality of available and potentially relevant data due to technological progresses in recording and storing information and it is driven by the diffusion of the internet and social networks. The consequence is the presence of a gigantic amount of data to manage and to analyze. Archives and data repositories of public administrations are incredibly rich "big data" sources too, although their effective exploitation in economic and statistical research has traditionally encountered several hurdles. As an instance, Big Data are often unstructured and noisy, so that preprocessing represents a fundamental part of the activity of the researcher. Indeed, the increasing complexity of economic phenomena, the high dimensionality of data and the need to combine textual and numeric information, make traditional approaches inadequate. Consequently, the potential for exploiting Machine Learning and Natural Language Processing (NLP) is limitless. NLP (or *text mining*) consists of techniques applying data mining concepts to textual data in order to extract quantitative information from qualitative ones. The object of the analysis is a *corpus*, that is a set of *documents*. A document in a database represents a statistical unit: in this thesis, documents are textual data referred to a firm, or official speeches released by central bankers. Text mining, coupled with learning algorithms, can be useful for error correction, missing data imputation, dimensionality reduction, topic modelling, information extraction and classification.

This work exploits Machine Learning and Text Mining tools addressing in an innovative way issues that have been widely investigated in the literature both at the micro and macro level. As far as the micro level is concerned, the object of interest are firms, demographic events characterizing them, their performances and the production of real-time official statistics. At the macro level, central banks communication is investigated, with a special focus on banking supervision. In fact, text mining allows to analyze with statistical tools texts referred to central bankers and policy makers. Whereas communication strategies aimed at monetary policies are deeply investigated, there is a lack of analysis on banking supervision, and this work partially fills this gap.

Machine Learning and Text Mining tools were applied to two kind of data: 1) textual data made publicly available online by the European Central Bank (ECB) and related to official speeches, interviews, press conferences and legal documents released by the ECB as an institution, or by the policymakers involved in ECB's boards and committees; 2) administrative data referred to firms recorded at the business register of the Chamber of Commerce of Bolzano. Those two data sources were exploited to address questions related to business demography and firms' performances at the

micro level, and central banks communication and banking supervision at the macro level.

This thesis was realized thanks to the technical and financial support of the Institute of Economic Research of the Chamber of Commerce of Bolzano. For that reason, each chapter was aimed at investigating statistical, economic and econometric techniques laying at the frontier of the research and capable of being applied at the local level. The common thread of the four submitted essays is represented by the use of Machine Learning, and especially Text Mining, with the following purposes: 1) data cleaning, preprocessing and missing data imputation; 2) converting qualitative information represented by textual data to quantitative ones capable of statistical analysis; 3) detecting hidden preferences from policymakers' statements; 4) providing real-time official statistics on business demography; 5) predicting firms outcomes, especially their survival probability. In those extents, this work is arranged as follows:

- The first chapter introduces the principal tools belonging to Machine Learning and Text Mining, providing instances of applications in several fields of the Economics. Some of the presented tools will be applied in chapters 2, 3 and 4 and the reader will be reminded to chapter 1 for a theoretical introduction.
- The second chapter is concerned with central banks communication strategies. Communication strategies adopted by central banks evolved during the last decade enlarging the toolkit available to central bankers. Nonetheless, communication has shown to be an extremely powerful unconventional instrument of monetary policy, especially at the zero-lower bound (ZLB). On this behalf, central banks can rely on communication in the extent in which they are perceived as credible and transparent; nonetheless, communication is effective as far as the conveyed message is clear and, especially during financial crises, unexpected. Due to the quantitative consequences that communication has on financial markets and economic outcomes, NLP emerged a technique to disentangle the effects that communication exerts in several domains and under different conditions. Nevertheless, NLP has been widely applied to investigate the role of communication for monetary policy effectiveness and financial stability related issues, but there is a lack of research on the role of communication in banking supervision. The second chapter fills this gap, providing a review of the literature applying text mining within central banks and to central banks, and estimating a Latent Dirichlet Allocation (LDA) using official statements of ECB's and SSM's members. Evidences are provided about the fact that, using the LDA, it was possible to predict the direction of SSM's agenda anticipating the emergence of Non-Performing Loans (NPLs) as a key topic already in 2014, at least two years in advance with respect to the inclusion of NPLs in official documents.

- Third and fourth chapters are extremely related themselves[1]. In chapter 3, an algorithm aimed at distinguishing apparently new activations linked to pre-existing firms from real start-ups is implemented. Missing data concerning the economic activity performed by firms (the NACE code) are imputed using textual descriptions provided by the entrepreneur after having trained a classification random forest. Moreover, a random forest is estimated to predict if an apparently death firm will be reactivated within two years.
    - The fourth chapter applies the algorithm presented in chapter 3 in order to distinguish new firms from reactivations. Moreover, firms' survival probabilities are estimated based on demographic and accounting data as well as real time information on the stance of the economy using a random forest trained on past observations. The accuracy of the predictive model was shown to depend crucially on the balancing among target classes in the training set.

We can conclude that NLP is extremely useful in investigating how institutions shape their strategies and the extent in which public debate and internal discussions affect the resulting policies and the following market's reactions. NLP presents the advantage of minimizing the extent of human discretionality in judging the content of a statement. If data scientists are concerned with the activity of collecting and manipulating data in order to make them capable of being analyzed, to economists the task of interpreting those massive amounts of information in order to investigate market's evolutions, consumers' confidence and, finally, economic agents' behavior is left. Machine Learning and NLP are shown to be extremely powerful tools for the researcher dealing with Microeconomics, Macroeconomics, Finance and in order to produce official statistics and to predict firms' performances.

[1] A draft version of the third chapter, "Classifying Firms with Text Mining", was presented at the poster session of the 29th (EC)^2 on Big Data Econometrics with Applications and published as a working paper by the DEM of the University of Trento. Moreover, the introduction and the literature review of "Predicting Start-ups Survival with Machine Learning" (section 1 and 2 of the fourth chapter) were written jointly by Giacomo Caterini and Matteo Cristofaro. Matteo Cristofaro is a researcher in Management and Industrial Organization at the University of Rome Tor Vergata. The data used in chapter 3 and 4 were partially provided by the IER of the Chamber of Commerce of Bolzano and are proprietary.

Machine Learning in Economics and Finance: A Review of the Literature

ABSTRACT

The role of Machine Learning techniques in Economics has increased rapidly in recent years. This is mainly due to the increasing amount of available data as a consequence of technological progress in collecting and storing them and, most of all, to the diffusion of the internet. As a matter of fact, the internet contains real time information capable of being exploited from an economic perspective. Once those data are stored, to economists is left the task of interpreting them in order to investigate and predict market's evolutions and agents' behavior. This is generally called Nowcasting. For this reason, in the last decade the use of alternative (or non-orthodox) data-sources, models and techniques aimed to study economical phenomena and the cooperation between economic departments and the computer sciences' ones have been encouraged and addressed. An overview of the applications of data and text mining techniques in economics and finance will be given. The role of text mining on predicting the effects of monetary policies will be stressed and further applications of such techniques in the field of forward guidance from central banks will be proposed. At the best of our knowledge, there is not a comprehensive review of the literature concerning data and text mining applications to monetary policy. This work also represents an attempt to fill this gap and to provide a new contribution to this branch of the literature.

Keywords: Machine Learning; Text Mining; Economics.

JEL classification: E00, C01, E40, E50

1 Introduction

The role of Machine Learning (ML) in Economics is becoming preeminent in the last decade. This is mainly due to the emergence of the *Big Data* (and a decisive contribution to such a phenomenon was given by the diffusion of the internet) and to the technological progress implying the availability of computers that are faster than in the past and equipped with higher computational power and larger storage ability. As Francis Diebold argues (Diebold, 2003), the expression "Big Data" should refer to the explosion in the quantity and quality of available and potentially relevant data due to technological progresses in recording and storing information. As a matter of fact, the internet contains real time hints about what economic agents are interested in, and that information can be stored and quantified. Professional agents use the internet in order to gather information about economic variables, statements from policy makers and shocking events, determining so a bidirectional exchange: on the one hand, messages are addressed via the internet and, eventually, the social media; on the other hand, what economic agents search for with their queries represents a new information itself. The consequence is the presence of a gigantic amount of data to manage and to analyze. If data scientists are concerned with the activity of collecting and manipulating data in order to make them capable of being analyzed, to economists the task of interpreting those massive amounts of information in order to investigate market's evolutions, consumers' confidence and, finally, economic agents' behavior is left.

In such extents, studying big data is strictly connected with machine learning techniques aimed to model complex relations through "*high-performance computer systems that can provide useful predictions*" overcoming computational constraints (Varian, 2014). Actually, big data is about collecting, managing and analyzing *real time* data relying on the fact that they can provide insights useful to predict *present* events more than future ones (Choi and Varian, 2012). For this reason, dealing with big data based on econometrics and machine learning techniques is often referred to as *Nowcasting*.

In the most recent economic literature, on the one hand the use of non-orthodox data-sources, models and techniques aimed to study economical phenomena, and on the other hand the possible cooperation between economic and computer sciences' departments were addressed and encouraged. To summarize, as Varian (2014) argues, a deeper understanding of the potentiality of applying ML tools to economics and econometrics is desirable since the recent increase in data availability, the consequent rise in the number of variables, with the need to select the most useful ones, and the need to overcome popular (over) simplifying schemes describing economic events according to linear models. As a matter of fact, most of the phenomena that we observe are so complex that we wouldn't believe.

In economic applications, “Machine Learning” usually refers to the use of Statistical Learning (SL) tools to store, manipulate and analyze big data (see Varian, 2014). This is probably due to the fact that, in classical economics and finance, the strongest (and most challenged) assumption concerns the rationality of economic agents, being the unbounded computational power of the human beings in their decision making process the natural consequence of it. In such extent, agents are supposed to take decisions analyzing the available amount of information like a machine would do. Machine Learning is actually a branch of Artificial Intelligence (AI). Such a discipline was born in the 50's with the ambition of developing computers which are able to think, in order to automate intellectual tasks normally performed by humans (Chollet, 2018). At the initial stages, classic AI (the so called *symbolic* AI) was mainly about programing the machine to execute a task following an already known set of procedures suggested by the researcher. Across the decades the discipline has evolved, pursuing the pioneering ambition by Alan Turing (1950) of achieving a computer capable of *learning* and *originality*[2], automatically finding the most efficient algorithmic procedure (being the efficiency evaluated according to computational time and the mismatch between the target and the effective result) and beyond the bounding of doing just what humans can tell it to do.

Roughly speaking, ML is concerned with the activity of learning from the observation of a specific set of data (sometimes called *examples*), and it is usually aimed to exploit the acquired knowledge in order to make predictions, often without the need to specify a detailed *a priori* statistical model (see Carbonell et al., 1983 and Samuel, 1959). According to Mitchell (1997) a computer is considered to learn from experience when its performance in addressing a specific task improves after the experience itself. As it will be shown, when the examples are labeled, we are in the realm of *supervised* learning. In the latter case, examples are represented by *constraints* over *attributes*: in simple words, the values that dependent and independent variables assume. In such a frame, *hypotheses* are specific sets of (joint) constraints on attributes and the learning activity consists in inferring the hypothesis best fitting the training examples (Mitchell, 1997). Although the border between them is rather thin, ML and SL share the same final aim of learning from data and are usually distinguished according to two aspects: the amount of assumptions concerning the model specification and the dimension[3] of data analyzed. It can be the case that learning models don't have closed form solution, so that iterative-algorithmic approaches and stopping rules for optimization processes are needed (see for instance LAR procedure for LASSO, Section 2.1.2). Even though both pursue the same task of learning, ML is usually concerned with developing computer algorithms to apply, as fast as possible, SL tools dealing with huge amounts of data, sometime specifying as less as

2 In his 1950's article, the British author focused on a digital computer replacing the humans in playing the imitation game.

3 As it will be shown, the problem of dimensionality concerns either the number of observations, features or both.

possible about the data generating process, sometime else involving researchers priors about studied phenomena (it is the case, as an instance, of Bayesian methods). Nevertheless, when dealing with ML from the theoretical point of view, the most quoted references are James et al. (2013) and Hastie et al. (2009): two popular books on Statistical Learning. As Chollet (2018) points out, ML basically is engineering oriented, meaning that inductive methods relaying on empirical observations play a central role more than mathematical and statistical theory. No matter the amount of initial assumptions, ML tools can span across the whole set of them without the need of a priori over-simplification and exploiting the computational power that follows the current level of technological process. Once the instances are observed, the aim of ML is assessing the underlying statistical structure allowing the machine to automate the task once a rule is identified. Our research poses the emphasis on *learning* and the differences between supervised and unsupervised learning will be pointed out, examining the main tools belonging to the different fields.

As pointed out by Athey (2018), machine learning basically deals with regression, classification and clustering. Simplifying, regression consists of estimating, starting from some labelled examples, a model predicting numeric values referred to some phenomena given a set of predictors. Classification is concerned with an analogous activity, but the output is categorical (classes). Finally, clustering consists of grouping observations according to their features in order to identify some similarities among them.

More in detail, Belloni et al. (2014) suggest exploiting the LASSO regression for variables selection in a causal inference framework. Regularization methods are aimed at prediction exploiting associations between the output *y* and a set of predictors (potential exogenous variables and controls) so that causal relations are difficult to advocate for, as well as the inferential validity of the estimated parameters. Furthermore, in the typical policy impact evaluation domain, variables that are correlated with the treatment (the ones causing endogeneity) would be shrunk toward zero by predictive algorithms based on regularization and dimensionality reduction, leading to omitted-variables bias (Belloni et al., 2014). Consequently, Belloni et al. (2014) suggest a "double selection" approach: when the researcher has to deal with thousands of possible control variables, a way to reduce dimensionality consists of selecting with LASSO a subset of controls affecting *y*, and running one more regression of the treatment on the whole set of control variables: in this way, the union non-zero coefficients obtained from the two separate models can be exploited to estimate the final inferential model assessing the effect of the treatment on the outcome. Similarly, when dealing with typical endogeneity concerns, the dimension of available instruments (eventually P>N, also considering transformations) can be reduced applying LASSO at the first stage of the 2SLS, in order to detect predictive instruments, and then using a standard econometric approach at the second stage, when the outcome is estimated on a subset of instruments selected at the first stage. More generally,

predictive models such as LASSO, RIDGE, regression trees and random forests can be applied to panel data to assess the effect of a treatment over treated by exploiting their forecasting power in order to predict effectively the outcome of the control group and to use it as a counterfactual for the analogue outcome of the targeted one after the treatment. Similarly, at the stage of policy design, heterogeneous effects of a treatment according to individual characteristics can be predicted to design assignment policies; tree-based methods can be particularly useful for those purposes, being the heterogeneity clearly represented by splits. Further applications are listed by Athey (2018), including the use of Factor Models to define structural models (reducing the dimension for the problem of the consumer's choice, as an instance) and for matrix completion when some entries are missing; as far as the latter aspect is concerned, using text mining to impute missing data in Chapter 2 should be seen in a similar fashion.

Motivated by the increasing role of learning economics and finance, several surveys provide an overview of the use of a wide range of tools for variegated applications. As an instance, Diebold (2012) is about the origins of the term "big data" and its use in the economic literature. Einav and Levin (2014) is an exhaustive review on the present role of big data in economics, suggesting how to improve economic research implementing new techniques belonging to other fields of study. The authors focus on the increasing availability of granular data and emphasize the potentiality of exploiting information coming from private sector and administrative records, stressing the role of technological progress in storing and processing data coming from such alternative sources, with a focus on the potential gains coming from combining ML and econometric tools. Varian (2014) provides some "Econometric tricks" and a basic introduction to ML tools such as classification and regression trees (CART), random forests (see also Biau and Scornet, 2016), bayesian structural time series (BSTS) and shrinkage methods as alternative to the traditional OLS estimator to compute regression coefficients in case of high dimensionality. Bholat (2015) is about the possibility of using big data for research in central banks with a special emphasis on the increasing interest that Bank of England is showing moving toward this direction. In Chakraborty and Joseph (2017) supervised and unsupervised ML techniques are exhaustively introduced and three case studies representing possible applications at central banks for banking supervision, inflation forecasting and clustering of high success financial technology firms are provided. Those were just few examples.

As far as text mining and NLP are concerned, Nassirtoussi et al. (2014) provide a review on the use of data mining and text mining (especially sentiment analysis) for the purpose of predicting a wide range of economic outputs, with a special focus on markets. Ravi and Ravi (2015) is mainly concerned with sentiment analysis. Kumar and Ravi (2016) briefly introduce the main data mining tools and text mining tasks presenting several applications in financial domain while Fisher et al. (2016) is a synthesis of the literature concerning NLP in accounting, auditing and finance. Regarding

the recent use of text mining for central banks, Bholat et al. (2015) is a quick overview of some text mining models with examples of applications in finance and monetary policy related studies.

Finally, at the best of our knowledge, there is not a comprehensive review of the literature concerning data and text mining applications to monetary policy, as it is generally meant. In fact, still the debate is ongoing and the interest on this field is quickly rising. The second chapter of this work also represents an attempt to fill this gap and to provide a new contribution to this branch of the literature.

This work is organized as follows: in the next Section data mining and text mining techniques are introduced, distinguishing between supervised and unsupervised learning. Some of those techniques will be applied in the other chapters composing the essay without further definition. In Section 3 a summary of the economic literature combining econometric techniques and data mining tools is provided.

2 Mining Techniques

The process of learning from data mainly occurs in two ways. Dealing with *supervised* learning, the researcher is provided two kind of variables: input variables (sometime called independent or explanatory variables, covariates, regressors, features or predictors) and output ones (dependent or response variables, target, regressed or predicted variables). Moreover, the number of features can be much higher than the number of explanatory variables, since features can be powers of the variables, interaction terms, transformations and so on. Variables can be *quantitative*, *qualitative* (sometime called discrete, categorical or factors) assuming a finite set of values and not being capable of ordering, or *ordered categorical* whereas categories can be ranked. In the easiest case in which there are just two categories, variables are coded via dummies. If outputs are not given, the researcher can be concerned with labelling variables, that is, matching to any input x_i one output y_i. Grouping (or clustering) observations according to common patterns (regularities in data discovered through computer algorithms, Bishop (2006)) whenever we have features without observing outcomes is the typical *unsupervised* learning problem. No matter how the labels of the outputs are obtained -manually or through clustering-, supervised learning is aimed to investigate the relation (if any) between inputs and outputs. Regularities among data are sought and once patterns are identified, ML is concerned with predicting the outputs from the observation of the inputs. Prediction tasks are: *regression*, when the output is a continuos variable and *classification*, whereas the output is qualitative (Bishop, 2006 and Hastie et al., 2009).

Another feature characterizing ML is the decomposition of the dataset in three subsets: the training, validation and testing sets. At the first stage the model parameters are estimated; at the second stage, the model is calibrated, that is, parameters are regularized involving in the estimation

procedure a tuning hyper-parameter, obtaining a balance for the trade-off between bias and variance; at the third stage, the model is tested in order to assess its out of sample validity. A peculiar purpose of ML is to obtain a model capable of being generalized (Blum et al., 2016). Estimating the model relaying on a single set we can be left, in the limit case, with a complex and flexible function exactly describing the data, with no bias but high variability. A drawback of such a procedure is that, once applied to a new dataset, the estimated model will perform very poorly in describing the data and predicting outcomes given the new inputs. This is due to the fact that the estimated model doesn't fit just the data but also the noise contained in them. A model precisely fitting the training data and failing clamorously when a new set is given is said to overfit the data: and that's why the sample should be divided in three parts (Hastie et al., 2009). Further details over the tradeoff between bias and variance will be provided later.

In the next section supervised learning techniques which have shown to be capable of application in the economic field will be briefly introduced; the next sections will be concerned with unsupervised learning, text mining and an exhaustive review of the economic literature coping with ML will be provided.

2.1.1 Supervised Learning: Regression

Suppose the real valued input vector$X \in R^P$, and the real valued random output $Y \in R$are jointly described by *P(X,Y)*. The typical learning problem is estimating a function of *X* predicting *Y*. This problem is usually addressed minimizing the expected value of a loss function *L(Y, f(X))* (Hastie et al., 2009):

$$E(Y - f(X))^2 = \int ([y - f(x)]^2 P(dx, dy)) \tag{2.1}$$

Being the loss function quadratic, the solution for the minimization problem is unique and the conditional expectation *E(Y|X=x)* is notoriously the function minimizing the expected loss (see Bishop (2006) for a formal proof and Wooldridge (2010) for further details). That is, *f(x)* is the regression function. Bearing in mind the Law of Large Numbers[4], let's place ourselves in a sample environment.

We have a set of *N* training data (x_1, y_1),..., (x_N, y_N) in which $x_i = (x_{i1}, x_{i2},..., x_{ip})^T$and we have to estimate the model (Hastie et al., 2009). We face the sample version of the problem (1.1):

4
Let's be x_1, x_2, … x_n *N* independent samples from a RV x, then: $Prob((((x_1 + x_2 + ... + x_n)/n - E(x))) \geq \varepsilon) \leq Var\,(x)/n\,\varepsilon^2$

$$f(x;\beta) = argminRSS(\beta) \tag{2.2}$$
$$= \sum_{i=1}^{N} ([y_i - f(x_i)]^2)$$

Estimating the model minimizing the *residual sum of squares* means *i*) estimating the coefficients vector β of the polynomial; *ii*) choosing the degree of the polynomial fitting the data. The simplest solution for (1.2) is, from Hastie et al. (2009), the linear regression model that represents, in a two dimensions space, the (unique) linear projection of *Y* on *X* (Wooldridge, 2010):

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_{j.} \tag{2.3}$$

where *P* stands for the number of regressors. The maximization problem (1.2) is solved by the OLS, imposing two restrictions:

$$\delta\, RSS/\delta\, \beta = -2X^T(y - X\beta) = 0 \tag{2.4}$$

$$\delta^2\, RSS/\delta\, \beta\delta\beta^T = 2X^T X = 0 \tag{2.5}$$

being $\boldsymbol{X}$ an *N x (p+1)* matrix (having as a first column a vector of ones for the intercept) and $\boldsymbol{y}$ an *N*-vector for which the pairs (x_i, y_i) are independent random draws and the error vector $\boldsymbol{U}$ is assumed to be uncorrelated with $\boldsymbol{X}$ from (1.4). If $(X^T X)$is nonsingular, the unique solution for (1.2) is:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{2.6}$$

Geometrically, the estimate $\hat{y} = X\hat{\beta}$is the orthogonal projection of $\boldsymbol{y}$ on the subspace of R^Nspanned by the column vectors of $\boldsymbol{X}$. Under (1.4) and (1.5), with variance of residuals independent from the inputs, the linear regression is the best linear predictor for the conditional expectation of *Y* given *X=x*, that is, the loss minimizer, and from the Gauss-Markow Theorem[5] follows that the OLS estimator $\hat{\beta}$is

[5] If: 1) X is a fixed *N* x (*P+1*) full column rank matrix, 2) $E(y) = X\beta$ and 3) $Var(y) = \sigma^2 I_n$, then $Var\beta - Var\hat{\beta} \geq 0\ \forall\hat{\beta}$linear in y

the best linear unbiased estimator for β(Wooldridge, 2015).

Considering a function that is linear in the parameters and posing ourselves, for the moment, in a comfortable cartesian plane with a single regressor *x*, the best representation of the data and, as a consequence, the best estimation of *f(x)* in (1.2) would be:

$$f(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M. \tag{2.7}$$

The linear regression is just a special case of (1.7). If the number of coefficients *M+1* equals the number of observations in the training data, then we obtain a polynomial with *M+1* degree of freedom perfect fitting the data. Unfortunately, as shown by Bishop (2006) and Hastie et al. (2009), the training error decreases with the degree of the polynomial but, after some threshold, the test error starts increasing dramatically. Nevertheless, the magnitude of the estimated coefficients increases with *M*, causing a severe variability in the model and poor out-of-sample predictive performances. Such phenomenon is known as *overfitting*.

In the world of big data, model's inputs are likely to be several, in the range of dozens, hundreds or even thousands. Datasets are actually described as *tall* (*N* big), *fat* (*P* big) or showing *PxN* huge. As the number of regressors increases, the amount of training observations needed grows dramatically. In particular, as argued by Bishop (2006), with *P* input variables a polynomial of order three would look like:

$$f(x,\beta) = \beta_0 + \sum_{i=1}^{P} \beta_i x_i + \sum_{i=1}^{P}\sum_{j=1}^{P} \beta_{ij}\, x_i x_j + \sum_{i=1}^{P}\sum_{j=1}^{P}\sum_{k=1}^{P} \beta_{ijk}\, x_i x_j x_k. \tag{2.8}$$

It means that for a polynomial of order *M* the number of coefficients grows at a rate of P^M. The phenomenon according to which the rise in the dimensionality implies a sparsity in the available data is known as *course of dimensionality* (Bellman, 1961). Furthermore, if the increasing complexity of the model reduces the bias, improving performances of the model in fitting the training set, a common drawback is represented by greater variability affecting the out-of-sample performance and lower interpretability. Assuming $Y = f(X) + \varepsilon$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$, Hastie et al. (2009) quantify the bias-variance trade off from the expected prediction error at $X = x_0$ (obtained comparing

and unbiased for β (Peracchi, 1995) .

the observed outputs and the ones predicted according to $\hat{f}(X)$) , based on the following decomposition:

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0] = \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$$

$$= IrreducibleError + Bias^2 + Variance. \quad (2.9)$$

It is clear that reducing the bias, the consequential increase in the variance due to greater complexity of the model will affect the prediction error and viceversa. For the linear model with *P* regressors, the expected prediction error is:

$$Err(x_0) = E[(Y - \hat{f}_p(x_0))^2 | X = x_0] = \sigma_\varepsilon^2 + Bias^2(\hat{f}_p(x_0)) + ||X(X^T X)^{-1} x_0||^2 \sigma_\varepsilon^2, \quad (2.10)$$

and the in-sample prediction error is:

$$\frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N} [f(x_i) - E(\hat{f}(x_i))]^2 + \sigma_\varepsilon^2 \frac{p}{N}. \quad (2.11)$$

It follows that the model complexity is clearly affected by the number of parameters *P*. Shrinkage methods aimed to reduce the model complexity are addressed in the next Section.

2.1.2 Shrinkage Methods

The underlaying idea for shrinkage methods is rather easy: a penalization term is imposed to the regression coefficients including a constraint in the minimization problem (1.2); this implies the regression coefficients to be shrunk toward zero.

Ridge regression was proposed by Hoerl and Kennard (1970) as the solution to the following minimization problem (Hastie et al., 2001):

$$\hat{\beta}^{ridge} = argmin_\beta \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \quad (2.12)$$

which reduces to the OLS problem when the penalization term $\lambda$is equal to zero. Intuitively, the higher the complexity parameter, the lower the magnitude of the coefficients, the lower the complexity and the consequent variance of the model. The following coefficients vector is obtained:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{2.13}$$

which clearly coincides with the OLS estimator but for the regularization term inside the parenthesis. Shrinkage is particularly useful in case of high correlation between variables, eventually causing positive magnitude of a parameter to balance negative magnitude of another one. Practically, multicollinearity of lines in $\mathbf{X}^T\mathbf{X}$ represents a threat for the invertibility. Hastie et al. (2009) provide further insights on the mechanism followed by Ridge regression showing the singular value decomposition (SVD) of the matrix **X** of inputs. Given an *N x p* matrix, its SVD is:

$$X = UDV^T = \sum_j \quad d_j u_j v_j^T \tag{2.14}$$

where ***U*** and ***V*** are *N x p* and *p x p* orthogonal matrices[6] having orthonormal columns vectors[7] $\boldsymbol{u_j}$ and $\boldsymbol{v_j}$ whereas ***D*** is a *p x p* diagonal matrix in which the entries $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$are *singular values* of ***X*** (Blum et al., 2016). If $d_j = 0$ for some *j*, ***X*** is singular. SVD provides the fit minimizing the perpendicular square distances between the data points and *k* dimensional subspaces, with $k<p$, that is, maximizing the lengths of the projections of the data points on the subspaces[8]. In other words, columns of ***V*** are the right-singular vectors of ***X***, being those vectors such that each $\boldsymbol{v_j}$ is orthogonal to $\boldsymbol{v_{j-1}}$ so that $\boldsymbol{v_j}$ is the vector starting from the origin and minimizing the squared (perpendicular) distance between the value points and the line parallel to the vector $\boldsymbol{v_{j-1}}$ or, analogously, it maximizes the length of the linear projection of the value point on the subspace spanned by $\boldsymbol{v_{j-1}}$. Each $\boldsymbol{v_j}$ is consequentially defined by the best fit line orthogonal to $\boldsymbol{v_{j-1}}$. Moreover, unlike the eigen decomposition, no special requirements are needed for ***X*** to perform SVD. Taking the fitted output vector and replacing the SVD of ***X*** in the coefficient estimator and in the input matrix we have what follows (Hastie et al., 2001):

---

6 $U^T = U^{-1}$and$V^T = V^{-1}$.

7 Two vectors are orthonormal if their length is equal to one and are orthogonal.

8 This simply follows from Pythagorean Theorem according to which the squared distance of each point to the origin is equal to the squared vertical distance between the point and its projection plus the squared length of the projection of the point on the subspace (Blum et al., 2016).

$$
\begin{aligned}
X\hat{\beta}^{ridge} &= X(X^TX + \lambda I)^{-1}X^Ty \\
&= UD(D^2 + \lambda I)^{-1}DU^Ty \\
&= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y.
\end{aligned} \tag{2.15}
$$

Clearly when d_j is zero ($\boldsymbol{X}$ is singular) the quantity $d_j^2/(d_j^2 + \lambda)$goes to zero too, shrinking toward zero the weight of the *j*-th inputs. Generally speaking, being the ratio$d_j^2/(d_j^2 + \lambda)$ smaller than one as far as $\lambda \neq 0$, columns *j* with d_j small are shrunk faster depending the speed of the penalization on the complexity parameter at the denominator. Corresponding small d_j to directions in the columns of $\boldsymbol{X}$ with small variance, those directions are shrunk the most; the rationale lays in the fact that, investigating causal relations between inputs and outputs, being the targets supposed to vary more reacting to variations on the regressors with higher variability, those ones are retained more with the hope to receive grater insights on the behavior of the response variable. In terms of *principal component analysis* (PCA), largest principal components are the directions maximizing the variance of the projected data. Being $v_j$the principal components directions of $\boldsymbol{X}$ and reflecting their ordering the magnitude of the sample variances of the linear combinations of $\boldsymbol{X}$'s columns, so that $z_1 = Xv_1$has the largest sample variance[9] among them, $z_1$is the first principal component of $\boldsymbol{X}$ and the next ones are smaller and orthogonal to the previous ones. It means that small d_j corresponds to smaller principal components ("lass principal") which are shrunk (Hastie et al., 2009).

The LASSO is a shrinkage method proposed by Tibshirani (1996). LASSO differs from Ridge for the shape of the constraint and, consequentially, the amount of shrinkage. The minimization problem is the following:

$$
\hat{\beta}^{lasso} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} (\beta_j)\right\}, \tag{2.16}
$$

because of the absolute value in the regularization component, the solutions are nonlinear in y_i and the estimation can be performed algorithmically by Least angle regression (LAR, see Efron et al., 2004 and Hastie et al., 2009). The different functional form of the constraints implies some

9 In this case the sample variance is $Var(z_1) = \frac{d_1^2}{N}$.

coefficients being exactly zero rather than simply shrunk toward zero, eventually leading to greater sparsity if LASSO is compared with Ridge. In such extent, LASSO and Ridge can be seen as Bayesian techniques with different priors (Hastie et al., 2009).

Shrinkage methods are usually applied for variable selection to cross-sectional data assumed to be independently distributed. In ML there is a rich debate about variable selection for time series. Konzen and Ziegelmann (2016) on the application of LASSO-type penalties in time series, Carrizosa et al. (2016) on VAR's sparsity and Kock and Callot (2015) on the "oracle property" and Demirer et al. (2018) applying the LASSO in time series in order to estimate financial networks are some examples.

2.1.3 Basis expansions

In daily life, and academic research as well, linear relations are rather unlikely to occur. As it was already stressed, linearity is actually a (very) special case of (2.7). Basis expansions of the input vector *X* are polynomial transformations replacing the inputs in the estimated (linear) model for the conditional expectation, that is, from Hastie et al. (2009):

(2.19)

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X),$$

where $h_m(X)$is the *m*th transformation of *X.* Although the transformation can take any functional form, the basis expansion of *X* is clearly linear in the coefficients so that simple least squares approaches are still valuable for estimation purposes.

2.2.1 Model Assessment and Selection: Cross Validation, Bagging and Bootstrapping

The idea underlaying to cross validation is rather simple. The dataset is divided in *k* parts, with *k* usually equal to 5, 10 or *N* (the latter case being *leave-one-out* cross validation), and the model is trained on the data contained in the *k-1* subsets and validated on the *k*th one. In formulas, from Hastie et al. (2009):

$$\hat{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k_i}(x_i)) \tag{2.17}$$

in which the fitted function $f^{-k}(x)$is computed on all the dataset but the *k*-th subset and loss function represents the prediction error as a function of the difference between the observations and the values predicted according to the training set. In order to set the penalization term, the CV criterion takes the form:

$$CV(\hat{f}, \lambda) = \frac{1}{N}\sum_{i=1}^{N} \quad L(y_i, \hat{f}^{-k_i}(x_i, \lambda)) \tag{2.18}$$

and $\hat{\lambda}$ is set at the value minimizing the test error curve.

2.2.2 Bootstrapping

The Bootstrap (Efron, 1982 and Efron and Tibshirani, 1994) is a resampling technique used to assess the extra-sample expected prediction error. According to the bootstrap, from the training set $Z = (z_{1,}z_{2,}\ldots, z_N)$, in which $z_i = (x_i, y_i)$, *B* equally sized independent subsets are randomly drawn with replacement. From each bootstrap dataset, estimates of population parameters are performed repeatedly, being *B* very big, in order to observe the behavior of the estimations across the *B* iterations. From Hastie et al. (2009), combining CV and bootstrapping we obtain the leave-one-out bootstrap estimate of the prediction error:

$$\hat{Err} = \frac{1}{N}\sum_{i=1}^{N} \quad \frac{1}{(C^{-i})}\sum_{b\in C^{-i}} \quad L(y_i, \hat{f}^{b}(x_i)), \tag{2.20}$$

where each $\hat{f}^{b}(x_i)$, for *b = 1, 2, ..., B ,* is the predicted output at $x_i$from the *b*-th bootstrap subset, and the sample error is averaged across the (C^{-i})samples not containing the *i*-th observation in order to avoid possible overlaps between training and validation set eventually overestimating the out-of-sample goodness.

2.2.3 Bagging

Bagging is a variation on a theme of the bootstrap. It consists in averaging predictions obtained based on *B* bootstrap samples in order to reduce the variability of learning methods. In formulas

(Hastie et al., 2009 and Breiman, 1996):

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x) \tag{2.21}$$

increasing the accuracy of the model as $B \rightarrow \infty$ and reducing its variance. For further details on how to divide observations in training and validation-test sets in econometrics and ML see Varian (2014).

2.3.1 Supervised Learning: Regression and Classification Trees

Regression Trees (Morgan and Sonquist, 1963, Breiman et al., 1984) are ML techniques aimed to deal with nonlinearities in the relation between inputs and outputs. The algorithm partitions the set of inputs in regions and a model is fitted in any of them. The algorithm works choosing at any step a variable and a split-point to divide the feature space; the model is fitted predicting the response variable as the average of the observed y_i in any region and the split-point is set to achieve the best fit; then, other split-points are chosen for the previously obtained regions and other variables are split until a stopping rule is reached. Consider the case of N (x_i, y_i)pairs, with $x_i = (x_{i1},\ldots x_{ip})$, that is p inputs and N observations, and suppose to divide the space of regressors in M regions R_1, ... R_M (Hastie et al., 2009) and represent the constant as c_m; then:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2.26}$$

from which it clearly follows that, if $f(x)$ is chosen minimizing $\sum(y_i - f(x_i))^2$, the best $\hat{c}_m$is the mean of $y_i$conditional to the region $R_m$to which the indicator function refers. Practically, starting from two regions R_1 and R_2, partitioning consists in finding the pairs (j, s), ie the predictor j and the split-point s, minimizing (from James et al., 2013):

$$\sum_{x_i \in R_1(j,s)} (y_i - \hat{y}_{R1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{y}_{R2})^2, \tag{2.27}$$

and repeating the process for each predictor and each cutpoint. As for any regression, we can increase the accuracy of the fit as much as we want simply increasing the number of regions in which the features are partitioned and, consequently, the number of terminal nodes. Intuitively, dividing the independent variable in an infinite number of small intervals and predicting the outcome with the

conditional average of the region is equivalent to perform a basis expansion or, more specifically, a local regression, in which the feature space is partitioned in sub-regions and some function is estimated for the response variable. When the phenomenon under analysis is likely to be nonlinear, trees perform better than linear regressions: even though regression trees work linearly via the average, partitioning the input space the researcher is able to capture nonlinear relations similarly to what happens with local linear regression. The extent in which the feature space is partitioned affects the bias-variance trade off since the size of the tree (namely, the number of branches) governs the complexity. Operationally, a large tree T_0 is estimated and then it is gradually *pruned* (that is, internal nodes are removed), reducing complexity similarly to what happens with the LASSO; at that stage, cross validation can be applied in order to select the tuning parameter based on the error rate in the test set. When the tuning parameter equals zero the obtained subtree coincides with T_0. Trees can also be compared, in some extent, to nearest neighborhoods, in light of the fact that in any region of the test set the observed response variables are predicted by a constant, namely the conditional mean of the observations included in the same region (that is, in the same neighborhood) of the training set. See Hastie et al. (2009) and Varian (2014) for further details and possible economic applications.

As the output of the supervised problem is a categorical one, the Classification Tree works analogously to the regression tree: the constant outcome predicted for a given partition of the predictors space of the test set is represented by the most frequent class in the same region of the training data and, estimating the tree, squared errors are replaced by entropy measures and the goodness of the tree is evaluated according to misclassification rates.

Trees are often characterized by high variability. In fact, if we estimate separate trees from two different training sets drawn from the same population, we can obtain extremely different results (Hastie et al., 2009). For this reason, the bagging procedure described in the previous Section can be applied to regression and classification trees, as it will be shown in the next chapters of this work. Briefly, several training sets are generated at random from the same initial sample using the bootstrap procedure, a tree is estimated for each sample and results are averaged (bagging) in order to obtain a smaller sample variance. Nevertheless, the resulting trees tend to be highly correlated in terms of variable importance measures as the same relevant variables will likely be detected at any iteration. For such a reason, Random Forests represent an improvement respect to bagged trees, picking at any split a subset of predictors chosen as candidates in order to exploit the variability resulting from estimating tress that are not too similar.

2.4.1 Machine Learning and Text Mining

By text mining (or *natural language processing*, NLP) is generally meant a set of unsupervised learning techniques applying data mining concepts to textual data in order to extract

quantitative information from qualitative ones (represented by character strings) and performing *information retrieval*, that is "finding material of an unstructured nature that satisfies an information need" (Manning et al., 2008). The information need can be seen as the topic the researcher is interested in. In such extent, text mining consists of content analysis. The object of the analysis is usually a *corpus*, that is a set of *documents*. A document in a database is a single observation, for instance a speech by a central bank governor, or a single statement by a member of a committee (see Manning et al., 2008, Hansen et al. 2017, Bholat et al., 2015). Formally, the document is generally represented as a list of words. According to Bholat et al. (2015) text mining techniques are distinguished by the epistemological approach as deductive and abductive ones. On the first extent, the focus is posed on Boolean[10] and dictionary text mining, in which a defined list of words is built up consistently with some existing theory. On the other hand, Descending Hierarchical Classification, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) start from the data in order to identify thematic patterns in the considered texts. Roughly speaking, the basic idea is that the frequency of words and their cooccurrence reflects some underlying structure to be assessed.

A common aspect of any text mining technique is that the analysis is preceded by a pre-processing phase in which the data -in that case, the documents- are prepared (as well as in econometrics or data mining). *Tokenization* consists of splitting a raw character in individual elements: in practice, sentences are divided according to the elements composing them (words, numbers, punctuation); moreover, for the purposes of the analysis some of those elements can be neglected reducing dimensionality without loosing relevant information: it is the case of *stopwords* like "a", "and", "is" and so on. The relevance of those words follows a negative power law (Zipf's law): even though stopwords are the most frequent ones in the corpus, nevertheless they don't reveal anything of interest for the purpose of the research (Manning et al. (2008)). Moreover words are mapped into their linguistic root eliminating affixes and retaining stems (from which it follows the term *stemming*); for instance, "gone" becomes "go" and terms like "unemployment"/ "unemployed", "growth"/"growing"/"increase"/"increasing" etcetera should be considered to belong to the same semantic group in the extent in which the underlying topic is the same. Finally, tokens are converted to lowercase, namely "US" becomes "us", "UK" becomes "uk" and so on. This pre-processing allows, for instance, more matches between the words present in the corpus and the words eventually listed in the predefined dictionary.

10
Boolean techniques won't be studied deeply in this work. Nevertheless, it is worth to say that they are based on the search for one or more key terms (Bholat et al., 2015) and according to them search engines provide results as the output of a query. Moreover, Baker et al. (2016) propose an uncertainty index based on Boolean techniques. The index is estimated counting the articles on the digital archives of 10 US's newspapers jointly containing terms related to three groups of words selected by the authors, and belonging to the semantic fields of economy, politics and uncertainty. It describes uncertainty on the extent and timing of monetary policy.

Texts can be represented in matrix form. Let's be $d \in \{1,2,\ldots, D\}$the set of documents; if each unique term in the corpus is indexed by $t \in \{1,2,\ldots, T\}$, than $\boldsymbol{M}$ is a D x T matrix with elements $m_{d,f}$ indicates the frequency of the t-th term in the d-th document. Such a matrix is typically known as document term matrix (DTM) and it is the tool allowing to map qualitative information (texts) to quantitative ones (frequencies). The DTM is the starting point for several machine learning application to texts, depending on the task: as an instance, as far as labelled texts are available to form a training set, we can estimate a model to classify documents based on words contained in them.

In the next Section a selection of text mining techniques will be briefly introduced. The emphasis on this work will be posed on dictionary methods, LSA and LDA being the most popular in economics in general and especially in monetary policy.

2.4.2 Dictionary Methods

Dictionary methods are based on constructing dictionaries, namely lists of words, considered to be relevant for the analyzed problem, meaning that the occurrence of a term in the document or the co-occurrence of a group of terms is considered to provide a signal for the existence of some underlying pattern, becoming the evidence stronger as the frequency[11] of the occurrences increases. Practically speaking, by "occurrences" is meant the match between a term present in the document and a term in the dictionary; terms are specified by the researcher choosing the ones more relevant for the specific purpose of the analysis. Moreover, terms are usually field specific, meaning that a specific dictionary should be prepared for any research question. Dictionary methods are widely used for many purposes: intuitively, once we group a set of words belonging to a specific field (for instance, in finance, the stock market) and we are able to assess the tone in which a specific topic is addressed, that is, we can distinguish between words involving optimism and pessimism, sentiment analysis can be performed. In such extent, an early example in finance is represented by Tetlock (2007) using Harvard IV-4 dictionary and principal component analysis (PCA) to provide evidences about high level of media pessimism being predictive of downward pressure on market prices and high trading volume.

11 Without bearing in mind that Zipf's law can represent a threat, considering high frequency of a word per se can be misleading. This possible drawback is addressed giving different weights to some words contained in different documents (see Manning et al., 2008 and Bholat et al., 2015).

2.4.3 Latent Semantic Analysis

The starting point of LSA (Deerwester et al., 1990) is decomposing the term-document matrix by SVD in order to obtain a *low-rank* approximation (Manning et al., 2008). Let's be $\boldsymbol{M}$ the D x T matrix previously introduced, then from (1.14):

$$M = UDV^T \tag{2.28}$$

(2.29)

and

$$M^T M = VDU^T UDV^T = VD^2V^T$$

In which the elements $\sigma_{ii} = \sqrt{\lambda_i}$of the diagonal matrix $\boldsymbol{D}$ are the singular values of $\boldsymbol{M}$ while D^2has entries the eigenvalues of $\boldsymbol{M}$ and $U^T U = I$follows from the orthogonality of matrix $\boldsymbol{U}$. It is worth noting that the rows and columns of the square matrix $M^T M$ correspond to each of the T terms and each (i, j)entry represents co-occurrences of words in documents. The elements of $M^T M$depend on the way $\boldsymbol{M}$ was originally expressed: if $\boldsymbol{M}$ simply reflects the occurrence of a term in a document through a dummy variable, then elements in $M^T M$provides the number of documents in which both *j*-th and *i*-th terms are present jointly (Manning et al., 2008).

After performing SVD we seek a low rank approximation for $\boldsymbol{M}$, that is, we obtain a matrix $M_k$having rank at least equal to *k* but much smaller that $\boldsymbol{M}$'s rank, under the constraint that *Frobenius* norm should be minimized, that is:

$$||N||_F = \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{T} N_{ij}^2}, where N = M - M_k \tag{2.30}$$

Which means, in words, that the lower rank matrix $M_k$should be as more as possible "similar" to $\boldsymbol{M}$. In such way, *r-k* singular values of $\boldsymbol{D}$ are truncated: this is equivalent to insert a weighting system shrinking them as shown in the previous Sections. At this point it clearly emerges the analogy between SVD in LSA and the constrained minimization in Ridge regression: finding a low rank matrix $M_k$we retain the (semantic) principal components of the text obtaining a new and simpler way to represent documents in the corpus. This is equivalent to consider just principal eigenvalues of the $\boldsymbol{M}$ matrix.

To conclude, as Manning et al. (2008) argue, LSA is aimed at dealing with *synonymy* and *polysemy* eventually underestimating measures of similarity[12] among documents after their vector space representation and its use in economics will be shown in the next Sections.

2.4.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) is a three-level generative probabilistic model. It is a Bayesian model in which items (documents in the earlier Sections) in a corpus are a mixture of underlying (latent) topics randomly drawn from a Dirichlet distribution assumed to be the prior of the model. Bayesian methods are based on few simple probabilistic rules. Considering two RV's X and Y for which $P(X, Y)\neq 0$, Bayes' theorem states that:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \tag{2.22}$$

Moving to the typical regression problem of estimating a parameters vector θ to explain the vector (or the matrix) of observed outputs Y, the (2.22) becomes:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{2.23}$$

in which the left hand side term is the *posterior* distribution of the parameter, representing a kind of improvement in our knowledge of θ after observing data, $P(Y|\theta)$ is the *likelihood function* for the observations $(y_1, y_2, \ldots, y_N)$ given θ describing the data generating process of Y given the set of parameters, $P(\theta)$ is known as the *prior* belief incorporating the amount of information the researcher owns about the phenomenon before looking at the data and $P(Y)$ is computed as

$$P(Y) = \int P\,(Y|\theta)P(\theta)d\theta. \tag{2.24}$$

The Bayesian approach consists in the assumption that we can learn about something unknown -the parameter θ in the mentioned case- from what we observe (Koop et al. (2007)). Since the interest lies in θ, terms like $P(Y)$ not containing it can be neglected and it can be written

12 Cosine similarity, for instance.

$$P(\theta|Y) \propto P(Y|\theta)P(\theta), \tag{2.24}$$

in words, the posterior probability of the parameters given the data is proportional to the product between the likelihood and the prior.

What distinguishes LDA from probabilistic LSA is the fact that a Dirichlet distribution is assumed as a prior for the topics, each word in a document can be assigned to several topics allowing the model to be more flexible and probabilities of documents and words over topics can be obtained. In words, each document contains a set of topics and each word listed in the document can be assigned to one or more topic. More formally, three steps are assumed for each document *d* described by a list of terms$w_d = (w_{1d}, w_{2d}, ..., w_{Nd})$in a corpus Ω: 1) the number *N* of unique words in a document is selected from a *RV* Poisson; 2) topics are distributed according to $\theta \sim Dir(\alpha)$, where the Dirichlet distribution for *k* topics has the following *pdf*:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}, \tag{2.31}$$

and 3) for each of the *N* terms describing the document $\boldsymbol{w}_d$ a topic z_{dn} is selected from θ_d and then the term itself is drawn according to the conditional distribution of terms over topics. To summarize, any word w_n in the list describing the document *d* is the result of a stochastic process in which a topic is drawn from a Dirichlet distribution and a word is extracted given the chosen topic, so that the joint probability of a list of *N* words and subset of *N* topics from the topic mixture θ is:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{2.32}$$

in which $p(w_n|z_n, \beta)$ is the probability of a given word to be drawn give the extracted topic z_n (Blei et al. (2003)), and conditional independence is assumed with respect to the latent parameters, that is, documents are independently but not identically distributed and at each slot the topic z_n is assigned independently (see the appendix of Hansen et al., 2017). Exploiting De Finetti's exchangeability theorem (De Finetti, 1970), Blei et al. (2003) end out with the marginal probability for a document to be composed by a given list of words for a specific set of topics:

$$p(w,z) = \int p\,(\theta)(\prod_{n=1}^{N} \quad p(z_n|\theta)p(w_n|z_n)). \tag{2.33}$$

To finalize, the main feature of LDA is identifying latent topics composing a document and providing the probability of a given word to occur in a certain topic. It follows that combining LDA and dictionary methods, once topics are assessed, according to the tone of words used and to the level of optimism or pessimism expressed by those words, sentiment analysis can be performed. In such extent, applications in economics and finance are provided in the next Sections.

3 Mining in Economics and Finance: Some Examples

According to Diebold (2012), the first academic paper in statistics or econometrics referring to Big Data was Diebold (2003), while the first significant academic reference is Weiss and Indurkhya (1998). In the last decade the interest in big data and ML and in the possibilities of application in the economic field is rapidly emerging. Specifically, the advantages of exploring those fields lay in the capability of exploiting ML techniques and the computational power of modern computers in order to manipulate and process big datasets, dealing with dimensionality (*N* x *P* is big), and in the possibility to collect and store real time data from non orthodox data sources for short run forecasting (also known as *nowcasting*).

This Section represents a (partial) review of the literature about the use of big data and machine learning in the economics. As far as big data are concerned, the focus is especially on the emergence of new data sources (especially posts on social media, queries on search engines and textual information scarped from the web); as to machine learning, the possibility to combine it with econometric tools to get more reliable insights from typical economic issues (predicting real and financial economic measures, providing official statistics, assessing the effect of an economic policy over the treated group and so on) is discussed, as well as the need to apply machine learning algorithms and new technologies to handle such a large amount of information.

As far as the computational and predictive point of view are concerned, there is a wide range of papers devoted to point out the possible applications of ML tools in econometrics for economics purposes. Varian (2014) encourages collaborations between economics and computer sciences department; the author suggests the implementation of ML tricks in econometric and presents statistical tools typically belonging to ML such as LASSO and Ridge regression, Random Forests, Classification and Regression Trees (CART), Bagging, Bootstrapping and Cross Validation, also providing examples in which they outperform standard econometric models. Belloni et al. (2014) on high dimensionality, treatment effect and using ML to make inference, Doornik and Hendry (2015)

about model selection with big data, Athey and Imbens (2017) on causality and policy evaluation are recent attempts to combine ML and Econometrics, as well as Mullainathan and Spiess (2017), and Fruet Dias and Kapetanios (2017).

Referring to nowcasting, economists and statisticians are asked to produce short, medium and long run predictions on economic variables like GDP, inflation or unemployment but estimates are often based on incomplete and uncertain data (see Castle et al. (2009)). Official data are available just many months after they are collected and processed. It follows the need to combine actual and known data and forecasts. Since the main indicators are usually available with a reporting lag and the estimates are revised moths later to obtain official results, reliable disaggregate data often are not available to estimate aggregate measures. Moreover, during crisis structural breaks are likely to occur and standard models are unreliable (see Castle et al. (2012)).

That was just a sample from the realm of reasons to use real time data to help economists computing short run predictions about the stance of economy. In such extent, in 2008 the Billion Price Project was created at MIT in order to obtain better real time estimates of inflation based on daily prices collected form hundreds of online retailers (see Cavallo and Rigobon, 2016). On the demand side, Redding and Weinstein (2016) use millions of bar codes associated with prices and sold quantities collected between 2004 and 2014 from Nielsen HomeScan database to build a unified price index and to estimate demand and elasticity of substitution. In the same spirit, Bajari et al. (2015) estimate 8 different models combining standard econometric techniques like conditional logit and panel regression and ML models such as bagging, random forests, support vector machine and shrinkage methods; they treat each prediction as a regressor for a new model in order to estimate demand curve based on a big dataset of scanner panel data from groceries[13].

The inclusion in econometric models of Google Trend time series as predictors is gaining popularity in economic research due to its the real-time informational content. Generally speaking, economic agents take decisions searching in advance and gathering information on the internet. As a matter of fact, the internet contains real time information about agents' decision process: for that reason the internet and social media are arising as feasible data sources capable of a wide range of applications. As a instance, a company named Expert System2 recently stated to be able to predict the outcome of the Italian constitutional referendum based on data collected on Facebook and Twitter through the Cogito software performing semantic analysis. Choi and Varian (2009) note a correlation between official data on "Initial Jobless Claims" in the US and Google Trend queries for key words related to job seeking and unemployment benefits. They add Google Trend time series to the baseline

13 The emphasize the potentiality of those techniques, just remind that in the 90's demand curve was estimated collecting few hundreds of price-quantity combinations at the fish market (Hardle & Kirman, 1995).

AR(1) model improving predictions. Choi and Varian (2012) used Google Trend data in order to predict automobile sales, unemployment claims, travel destinations and consumer confidence. The authors improve the estimates of motor vehicles and parts sales adding to an AR(1) + seasonality process the variable represented by the Google Trends index. In the analyzed case, adopting the time series representing queries for words "suv" and "insurance" a considerable improvement in the *out of sample* forecast is obtained. In the cases proposed by Varian, adding a new regressor taken from the internet (in this case the index realized by Google Trends indicating the interest in the words "suv" and "insurance") improves the regression reducing the SE and the mean absolute error (MAE) of the forecast. Intuitively, the idea is that agents planning to buy a car start to look for insurance companies or compare models on the internet. In the case in analysis, the advantage of such AR(1) + seasonality + Google Trend process is represented by the fact that an estimate of the amount of sales is obtained few weeks before official data are issued. Using the Google Trend index as a regressor approximating the interest in automobile, Carrière and Labbé (2013) predicted automobile sales in Chile; the authors also pointed out that requests for identical queries on different days return different series, suggesting that sampling takes place any 24 hours: for this reason the researcher s encouraged to download the time series several times and using the sample mean as a predictor. Donadelli (2014) measures policy-related uncertainty through Google searches for "stock market", "Fed" and "US politics" finding out that higher uncertainty so measured is predictive of drop in prices of quoted assets and reduction of industrial production. Siliverstovos and Wochner (2018) estimate touristic demand in Switzerland based on Google Trend time series. Nevertheless, Lazer et al. (2014) concerns possible traps of using Google Trend highlighting the poor performances of Google Flu Trend in estimating influenza. On the role of Google Trend for nowcasting see also Shimshoni et al. (2009).

Varian (2014) provides a basic introduction to Machine Learning techniques like classification and regression trees (CART), random forests (see also Biau and Scornet (2016)), bayesian structural time series (BSTS) and the LASSO technique, a shrinkage method used as an alternative to the OLS estimator in order to compute the regression coefficients in case of several (in some case, hundreds) regressors. An example of CART provided in Varian (2014) concerns predicting the number of survived people on the Titanic based on three regressors: *classes*, *age* and *sex.* Varian shows that some phenomena are too complex to be described by linear models. Standard approaches in analyzing the Titanic dataset would neglect the role of the nonlinearity in the *age*. The CTREE above is just an example and the observations in the dataset are few thousands: at the bottom of big data and machine learning there is the idea to analyze millions of data. Further advantages of combining econometrics and machine learning in the economic analysis are pointed out in Varian (2016) in which Experiments, Regression Discontinuity, Instrumental Variables and Difference in Differences are dealt with and insights about the use of inferences for Marketing are provided.

Other applications in the literature are Cooper et al. (2005) for cancer related topics, Preis et al. (2010) for consumer sentiment, Gruhl et al. (2005), Antenucci et al. (2014) and Asur and Huberman (2010) for predictions using web searches and social media like Twitter. Further references on Bayesian techniques and Machine Learning are Hastie et al. (2009) and Scott and Varian (2014).

Earlier sections were devoted to supervised and unsupervised learning, with a special emphasis on text mining. Even though the focus of the present and of the next chapter is on the role of text mining in monetary policy, still it is useful to present quickly some applications outside that nest, in economics and finance. A pioneer paper in such extent was Tetlock (2007), using Harvard IV-4 dictionary and principal component analysis (PCA) to provide evidences about high level of media pessimism being predictive of downward pressure on market prices and high trading volume. Bollen et al. (2011) use OpinionFinder and GPOMS to assess Twitter mood in order to predict the stock market evolutions. Renault (2017) performs text analysis over StockTwits platform to investigate investors sentiment. Renault constructs a lexicon of words used by investors sharing opinions and emphasizing the importance of a dictionary strictly field specific. Then, comparing estimated sentiment and S&P 500 index he points out that the first 30 mins in which the mood changes can be predictive of the market movements. Finally, Bandiera et al. (2017) exploits LDA to investigate the relation between CEO behavior and firms' performances.

4 Conclusions

In this chapter the main supervised and unsupervised ML techniques were described, distinguishing between data mining and text mining, and an accurate description of their use in the economic literature was provided. The emphasis was mainly posed on the application of machine learning in economics and finance.

In next chapter, differences between ECB, BoE and FED will be described with regards to transparency and communications strategies, discussing textual data availability. NLP will be applied to official texts available on the website of the European Central Bank to investigate the way in which European Single Supervisory Mechanism (SSM) conducts its decisional process, the role of potential influencers on the evolution of the debate and the extent in which final guidelines reflects such influencing debate.

Since at the best of our knowledge, there is not a comprehensive review of the literature concerning data and text mining applications to monetary policy, as it is generally meant, next chapter also represents an attempt to fill this gap and to provide a new contribution to this branch of the literature.

References

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using social media to measure labor market flows* (No. w20010). National Bureau of Economic Research.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492- 499). IEEE.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3-32.

Athey, S. (2018). The impact of machine learning on economics. In The economics of artificial intelligence: An agenda. University of Chicago Press.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5), 481-85.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636.

Bandiera, O., Hansen, S., Prat, A., & Sadun, R. (2017). CEO Behavior and Firm Performance (No. w23248). National Bureau of Economic Research.

Bellman, R. E. (2015). Adaptive control processes: a guided tour. *Princeton university press*.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, 28(2), 29-50.

Bholat, D. (2015). Big data and central banks. *Big Data & Society*, *2*(1), 2053951715579469.

Bholat, D. M., Hansen, S., Santos, P. M., & Schonhardt-Bailey, C. (2015). Text mining for central banks.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Blum, A., Hopcroft, J., & Kannan, R. (2016). Foundations of data science. *Vorabversion eines Lehrbuchs*.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. *CRC press*.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, *32*(4), 289-298.

Carrizosa, E., Olivares-Nadal, A. V., & Ramírez-Cobo, P. (2016). A sparsity-controlled vector autoregressive model. *Biostatistics*, *18*(2), 244-259.

Castle, J. L., Fawcett, N. W., & Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, *210*(1), 71-89.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2012). Model selection when there are multiple breaks. *Journal of Econometrics*, *169*(2), 239-246.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. *The Journal of Economic Perspectives*,*30*(2), 151-178.

Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. BoE Staff working paper N°674.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends.*Economic Record*, *88*(s1), 2-9.

Chollet, F. (2018). *Deep Learning with Python*. Greenwich, CT: Manning Publications CO.

De Finetti, B. (1970). Teoria della probabilità. Einaudi.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Demirer, M., Diebold, F. X., Liu, L., & Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1), 1-15.

Diebold, F. X. (2003, January). 'Big Data'Dynamic factor models for macroeconomic measurement and forecasting. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,"(edited by M. Dewatripont, LP Hansen and S. Turnovsky)*(pp. 115-122).

Diebold, F. X. (2012). On the Origin (s) and Development of the Term 'Big Data'.

Donadelli, M. (2015). Google search-based metrics, policy-related uncertainty and macroeconomic conditions. *Applied Economics Letters*, *22*(10), 801-807.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Society for industrial and applied mathematics.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*,*346*(6210), 1243089.

Fruet Dias, G., & Kapetanios, G. (2017). Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets. *Journal of Econometrics*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer, New York, NY.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National academy of sciences*, *107*(41), 17486-17490.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005, August). The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 78-87). ACM.

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2), 801-870.

Hardle & Kirman (1995). Nonclassical demand: A model-free examination of price-quantity relations in the Marseille fish market. *Journal of Econometrics, 67(1), 227-257.*

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, *12*(1), 69-82.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: Springer.

Kock, A. B., & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2), 325-344.

Konzen, E., & Ziegelmann, F. A. (2016). LASSO-Type Penalties for Covariate Selection and Forecasting in Time Series. *Journal of Forecasting*, *35*(7), 592-612.

Koop, G., Poirier, D. J., & Tobias, J. L. (2007). Bayesian econometric methods. *Cambridge University Press*.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-*

Based Systems, 114, 128-147.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval, p.1.

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37), 870-877.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302), 415-434.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

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

Peracchi, F. (1995). Econometria. McGraw-Hill libri Italia.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14-46.

Redding, S. J., & Weinstein, D. E. (2016). A unified approach to estimating demand and welfare.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance*, 84, 25-40.

Scharfstein, D. S., & Stein, J. C. (1990). Herd behavior and investment. *The American Economic Review*, 465-479.

Scott, S. L., & Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1-2), 4-23.

Shimshoni, Y., Efron, N., & Matias, Y. (2009). On the predictability of search trends. *Google Research Blog*.

Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization*, 145, 1-23.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3- 27.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, *113*(27), 7310-7315.

Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Nelson Education.

Predicting NPLs with NLP

ABSTRACT

Communication strategies adopted by the most influencing central bankers are evolving in the last decade. This work provides a comprehensive review of the literature concerning data and text mining applications to monetary policy. Moreover, we suggest applying text mining to the analysis of official documents released by banking authorities. An overview of the applications of machine learning and text mining techniques in investigating central banks' communication strategies will be given. The role of text mining on predicting the effects of monetary policies will be stressed and further applications of such techniques will be proposed. Indeed, we apply the Latent Dirichlet Allocation to ECB's and SSM's textual data released between 2014 and 2017. We show that, already in 2014, it was possible to predict the emergence of a core topic, that is banks' Non-Performing Loans, officially appearing in the SSM's political agenda just in the late 2016.

Keywords: SSM; ECB; Communication; LDA.

JEL classification: E00, C01, E40, E50

1 Introduction

Communication strategies adopted by the most influencing central bankers are evolving in the last decade. Studies on central banks' communication heavily addressed its adoption as an unconventional tool pursuing monetary and financial targets: as an instance, modelling long-term interest rate expectations[14]. Over the years, central bankers shifted from delivering parsimonious and obscure messages, to a rapid increase in the degree of transparency and clarity. This is mainly due to the growing consensus around the consideration of central banks communication as a tool to influence expectations and to reduce uncertainty (see for instance Woodford, 2005 and Campbell et. al., 2012, while Blinder et. al., 2008 provide a comprehensive survey). The European Central Bank (ECB) started providing forward guidance since July 2013 (see ECB, 2014). Praet (2013) defines forward guidance as "a communication instrument by which central banks convey their monetary policy orientation going forward, conditional on their assessment of the economic outlook". As a matter of fact, central banks influence the economy modifying short term interest rates - affecting in this way the long-term structure of interest rates - and providing signals about future policies based on forecasts. Moessner et al. (2017) provide a survey on the effects of communication on the economic and financial variables, comparing theory and empirical evidences. Those authors distinguish between Delphic communication (involving forecasts and describing likely monetary policies) and Odyssean communication (indicating central bank's commitment to a specific policy) and operate a further distinction, introducing the definition of Aesopian forward guidance for communication about policies adopted under specific (and unusual) scenarios. According to Moessner et al. (2017), no central bank implements Odyssean communication in practice; moreover, they consider ECB's communication to be "open ended", that is, no specific information are given on the timing and magnitude of monetary policies, nor on their conditionality to underlying macroeconomic stances. Nevertheless, at least in recent years, ECB's communication has moved toward a greater adoption of time and state contingent forward guidance. The speech released by Mario Drahi on March 2019 is an example:

> "[…] we decided to keep the key ECB interest rates unchanged. We now expect them to remain at their present levels at least through the end of 2019, and in any case for as long as necessary to ensure the continued sustained convergence of inflation to levels that are below, but close to, 2% over the medium term."[15]

Such a statement doesn't represent a binding commitment but, at least, a form of forward guidance

14 See Bini-Smaghi (2009) on the definition of unconventional policies and on the set of possible strategies a policy maker can implement.

15 Draghi (2019), available at: https://www.ecb.europa.eu/press/pressconf/2019/html/ecb.is190307~de1fdbd0b0.en.html

far from being "open ended".

The theoretical and empirical literature on central banks communication for monetary policy purposes and financial stability is extremely rich and well established. Born et al. (2014) contributed decisively to shed light on the effects of communication on financial stability performing a textual analysis over central banks Financial Stability Reports (FSRs), official speeches and interviews released by policymakers. According to the authors, central banks communication can be aimed at pursuing financial stability or at enhancing the effectiveness of monetary policy. In turn, communication is an unconventional tool for monetary policy. Independently on the purpose, communication effectiveness depends crucially on central bank credibility and on the clarity of the conveyed message. Nevertheless, there is lack of analysis on the role of communication on micro-prudential supervision. As a matter of fact, Natural Language Processing (NLP)[16] has been widely applied studying unconventional monetary policies, but communication related to banking supervision has received less attention. In fact, the article by Carretta et al. (2014) represents the principal application of textual analysis to assess different supervisory styles among supervising authorities, but no further contributions apply NLP to the European banking supervision[17]. The present article provides a new contribution to the literature investigating the role of communication for the European Single Supervisory Mechanism (SSM) and retrieving hidden clues in official statements.

After the 2008's financial crisis, policymakers devoted their efforts at the implementation of microprudential measures aimed at enhancing macroeconomic stability. Since 2014, European banking supervision was delegated to the Single Supervisory Mechanism (SSM), and non-performing loans (NPLs) arose as a central topic dominating the political debate and raising the academic interest (see for instance Gaffeo and Mazzocchi, 2019). According to ECB (2017a, b), low asset quality issues related to high levels of NPLs affect liquidity and make the recovery slower, reducing banks' profitability. Nonetheless, similarly to asymmetric effects of exogenous shocks, the heterogenous features characterizing financial players in the European banking union imply that changes to the common supervisory framework are perceived non-homogeneously. Consistently with Schmeling and Wagner (2019) -showing that ECB tone changes in the press conferences after Governing Council meetings are predictive of future changes in monetary policy rates-, we believe that quantitative insights can be gained from the statistical analysis of official statements from ECB's SSM supervisors, as those represent the inputs determining final supervisory decisions. As a matter of fact,

16 By NLP is generally meant the application of machine learning to texts in order to extract quantitative information from qualitative ones.

17 Instead, Goldsmith-Pinkham et al. (2016) deal with Fed's banking supervision.

the increasing complexity of economic phenomena, the high dimensionality of data and the need to combine textual and numeric information, make traditional approaches inadequate. Consequently, the role of Machine Learning (ML) and NLP is becoming preeminent for a wide range of economic applications (see Varian, 2014, Athey and Imbens, 2017, Mullainathan and Spiess, 2017 and Athey, 2018). NLP is proving to be extremely powerful in investigating how institutions shape their strategies and the extent in which public debate and internal discussions affect the resulting policies and the following market's reactions. Each document in a corpus refers to a statistical unit: for instance, a speech by the governor of a central bank. Because of their massive scale and peculiar features, textual data cannot be handled by humans in a fast and costless manner. Moreover, NLP presents the advantage of minimizing the extent of human discretionality in judging the content of a statement. As emphasized by BoE (2015) and Chakraborty and Joseph (2017), there is potential for dramatic gains by exploiting NLP to address communication and decision making-related topics, especially for central banks.

In this article the LDA, a model of unsupervised learning, is applied to official statements of ECB and SSM members. LDA is a clustering algorithm proposed by Blei et al. (2003) assuming each text to be composed by a mixture of latent topics. We provide evidence about the fact that, analyzing texts with topic modelling, we were able to anticipate the emergence of key topics addressed by the supervisory authority and to identify which country in the supervisory board contributed the most to dictate the political agenda.

Whereas Moro et al. (2014) exploit LDA to perform a literature analysis from 2002 to 2013, Einav and Levin (2014) is an exhaustive review on the role of big data in economics, Nassirtoussi et al. (2014) provide a review on the use of data mining and text mining for the purpose of predicting economic and financial outputs, Kumar and Ravi (2016) briefly introduce the main data mining tools and present several applications in financial domain and Fisher et al. (2016) is a synthesis of the literature concerning NLP in accounting, auditing and finance, at the best of our knowledge there is not a comprehensive review of the literature concerning Natural Language Processing (or text mining, TM) applications to central banks communication. In fact, the debate is still ongoing and the interest on this field is quickly rising. This work represents an attempt to fill this gap, taking the stock of what has been done so far in applying ML and NLP to study central banks communication.

The rest of the paper is organized as follows: in section 2 we summarize the existing works in which NLP tools are applied to central banks documents; in section 3 we briefly discuss the organizational structures of ECB and SSM, dwelling on data availability; in section 4 we apply LDA to ECB's and SSM's texts; section 5 further discusses the main conclusions of this survey.

2 Text Mining Techniques for Central Banks Communication: A Review of the Literature

Communication strategies adopted by central banks are evolving and the most influencing central bankers are moving toward greater transparency and clarity. In fact, there is a growing consensus around the consideration of central banks communication as a tool to shape expectations and to reduce uncertainty, as far as monetary authorities are credible and assumed to possess asymmetric information (see Woodford, 2005, Ehrmann and Fratzscher, 2007, Campbell et. al., 2012 and Moessner et al., 2017). Nonetheless, according to Bulíř et al. (2013) whom analyze the complexity of seven central banks official documents and released statements, clarity of communication is country and institution-specific being also partially affected by the economic contest and by the institutional decision process, although their evidences are weak.

Three dimensions in which communication matters can be distinguished: *i*) monetary policy, *ii*) financial stability and *iii*) banking supervision. Furthermore, among communication concerning monetary policies and financial stability, we can distinguish between conventional statements (the bare communication of the policy decision) or unconventional ones (the central banker depicts possible scenarios for interest rates, inflation and financial markets outcomes). As far as the role of machine learning is concerned, we differentiate between the adoption of ML tools *within* central banks and *outside* central banks, among researchers and professionals. In next Sections we summarize the direction that central banks' research has taken, how departments are equipping themselves, and the main contributions in quantifying the role of communication on economic, financial and supervisory outcomes.

As far as the general role of central banks communication is concerned, according to Blinder et. al. (2008), communication "creates" news and guides agents in forecasting the likely evolution of economic phenomena and monetary policies. As a consequence, forward guidance is considered as a powerful unconventional tool for monetary policy, especially at the zero-lower bound (see Wu and Xia, 2016). With regard to financial stability, Born et al. (2014) stress the importance of official speeches and interviews in providing unexpected information, since official reports are often precisely scheduled. The authors perform a textual analysis on the content of FSRs and of governors' speeches and interviews using the software Diction 5.0 to scores texts along the optimism dimension in order to assess the central banker's sentiment. The authors pointed out that, in normal time, optimistic FSRs are effective in reducing uncertainty (that is, returns volatility of financial assets) and in improving stock market sentiment, while during financial crises non-scheduled statements are more effective as they "surprise" the markets, proving to be flexible tools. The latter result is consistent with findings by Rosa and Verga (2007) explaining innovations in market expectations with

unexpected information released by the ECB. Carretta et al. (2014) investigate in their pioneering article the effectiveness of different styles of banking supervision. They analyze official statements made by the heads of several authorities using a vocabulary of words identifying six types of national cultures to assess which kind of supervision is more likely to reduce the risks of default. The authors find that a supervisory culture oriented to collectivism (that is, oriented to reduce uncertainty and enhancing overall stability of the system) increases the banks' distance to default, while "power distance" oriented culture decreases the banks' distance to default, being less flexible and more authoritative.

Once we have summarized how communication matters in monetary policy, financial stability and banking supervision, in the next Sections we take stock of the literature on the role of communication in central banks and national supervisors for different purposes, with a special emphasis on the use of text mining tools. We highlight how NLP and ML are exploited by central banks, to assess the efficacy of their own communication strategies, and by professionals, in analyzing official documents referring to central bankers and supervisory authorities. In such extent, as we will show, since the American Fed is the most transparent among central banks – that is, the one for which the richest collection of documents is publicly available - the literature on communication and decision making in which NLP is applied to the US is well established. As far as the Europe is concerned, the Bank of England started to emphasize the potential of ML rather early, providing seminal contributions. Finally, we will highlight the way in which ECB's approach to communication is evolving.

2.1.1 Machine Learning at the Bank of England

The Bank of England (BoE) has been pioneering in several domains of central banking. As an instance, according to Born et al. (2014) BoE was the first central bank introducing FSRs in 1996. Putting in relation the tone of BoE's FSRs (as estimated by the software Diction 5.0) and financial markets outcomes, the authors also point out that communications on financial stability released by the BoE exert international spillovers reducing volatility.

Among the principal central banks, BoE started investigating the potential for applying text mining relatively early. In 2014, BoE hosted an event called "Big Data and Central Banks", concerning the increasing importance of big data and machine learning for central banks in the present and in forthcoming years. Most of all, the emphasis was on the increasing availability of granular data and on the possibility of using them for economic research (Bholat, 2015). In its "One Bank Research Agenda" (OBRA), BoE (2015) mentions the use of big data, including the use of internet and social media as new data sources, and the application of ML techniques, including NLP, to better

understanding economic and financial systems, and to assess financial markets' and consumers' sentiment evolutions. Bholat et al. (2015) provide an overview of text mining methods such as Boolean and dictionary techniques, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Descending Hierarchical Classification; moreover, they highlight the possible advantages of central banks exploiting text mining on economics and finance. For instance, Nyman et al. (2018) build a sentiment index representing the balance between excitement and anxiety in the financial market. They analyze texts taken from three sources specifically related to the financial system (BoE daily market commentary, broker research reports and Reuter's news articles in the UK) defining a dictionary of relevant words according to their theoretical frame (deductive approach), that is the idea that agents corroborate their expectations on gains or losses constructing narratives. If comparing different data sources are observed homogeneity of beliefs among markets participants and similar emotional contents, it can be inferred the direction of consequences likely to hit the financial market. In such extent, the authors argue that shifts in sentiment, as detected by text mining techniques, are predictive of increasing volatility and structural breaks in financial markets and can anticipate the increasing anxiety immediately preceding financial events such as Lehman Brothers collapse and the 2011-2013 European sovereign debt crises[18]. Their index also leads the trend of other uncertainty indexes, such as the one proposed by Baker et al. (2016): that index measures policy-related uncertainty estimated counting the articles on the digital archives of 10 US's newspapers jointly containing terms related to three groups of words selected by the authors, and belonging to the semantic fields of economy, politics and uncertainty. The index proposed by Baker and his co-authors is well known in the literature[19] and represents an instance of Boolean technique; nevertheless, exploiting the gigantic potentiality of text mining, accurate analysis can be performed and computational time can be reduced.

In his recent work, Joseph (2019) investigates the trade-off existing in ML between models' predictive accuracy and statistical interpretability. In fact, models such as support vector machine, random forest and artificial neural networks are often referred to as "black boxes" because of their poor suitability for inferential purposes that, viceversa, dominate econometric approaches. Chakraborty and Joseph (2017) exhaustively introduce supervised and unsupervised ML techniques and provide three case studies representing possible applications for central banks, that is: *i*) automated inspection of balance sheets on behalf of banking supervision, *ii*) inflation forecasting combining standard econometric models and machine learning ones such as Ridge regression and

18
A possible drawback of such an approach is the following: in a system in which all the financial participants use text mining for sentiment analysis, as increasing anxiety is detected agents could decide simultaneously to close their long positions, so causing the market to fall as in a self-fulfilling prophecy.

19
Their article was published by QjoE and the 16/12/2017 it had 1495 citations on Google Scholar.

SVM and *iii*) clustering of successful financial technology firms. Focusing on banking supervision, the authors train several models aimed at predicting the occurrence of alerts from supervisory authorities based on accounting[20] items raising concerns. The most powerful model (that is, the one with highest predictive accuracy) is found to be the random forest and the bank's profitability is ranked as the main feature showing the highest predictive power and ability to reduce the classification error at any iteration.

Bholat et al. (2017) compare style of confidential Periodic Summary Meeting (PSM) letters sent by two different UK's supervisors-FSA and PRA- to banks, after FSA was disbanded because of inadequate supervision. The authors want to assess if supervising style changed over time and if there is a proportionality in dealing with different institutions accordingly to their specific exposure to risk and systematic possible impact, expressed by ranks. 25 linguistic features related to five high level groups (measures of linguistic complexity, sentiment indicators, directiveness, formality and forward-lookingness) are constructed. As the dataset is wide (the number of observations N is small and the number of regressors P is large), the authors exploit random forests to classify the category of the firm (that is, the capacity to cause damages to the UK financial system) according to the tone of the letter. Moreover, random forests are very powerful in handling non linearities and interactions[21]. The authors conclude that PRA's letters are proportional to the idiosyncratic and systemic risk of the firm. Letters to highly risky firms are more detailed, less directive and more assertive. Nevertheless, PRA's letters are clearer and more detailed than FSA's, focusing on liquidity requirements and presenting the perspective of possible default.

2.1.2 The Fed's Forward Guidance and the FOMC Meetings

Across the years Federal Reserve (Fed), the American central bank, gradually increased its effort aimed at transparency benefitting financial markets anchoring inflation expectations (see Hernández-Murillo and Shell, 2014). Since 1993, the Fed started publishing meetings' transcripts of Federal Open Market Committee (FOMC) -the body within the Fed in charge of setting monetary policy- with a five-year lag and minutes after six weeks. The first postmeeting statement following the FOMC's monetary decision was released in 1994 and since 1999 the FOMC started releasing a statement after any meeting, independently on the decision. Moreover, in 2002 the FOMC started providing immediately after the meeting voting information and a short explanation of any dissenting votes. Meade and Acosta (2015) observe the time evolution of the informative content of postmeeting

20 For a comprehensive synthesis of the applications of NLP in accounting, auditing and finance see Fisher et al. (2016).

21 In linguistic field, by "non-linearities" the fact that negligible changes in the syntax imply dramatic changes in the meaning is meant.

statements and of their length. The authors measure the correlation of words used in two consecutive statements estimating cosine similarities between speeches and assessing how the degree of standardization of statements evolves over time. They find an overall increasing level of semantic similarity characterized by a pronounced volatility, especially during the 2008's financial crises, due to the need faced by policymakers to react to unexpected evolutions of the stance of the economy. Consistently, the complexity of statements has increased with their length, especially during periods of unconventional monetary policies (Hernández-Murillo and Shell, 2014).

Kahveci and Obadas (2016) describe the degree of certainty and optimism in Fed's postmeeting statements, finding out that after the 2007's financial crises the Fed reduced its optimism in statements and increased the level of certainty reflecting forward guidance.

Central banks' efforts are aimed at enhancing their credibility through transparency, being credibility the *condition sine qua non* for policies effectiveness. Hayo and Neuenkirch (2015) assess the role of major central banks' communication on their perceived credibility, independence and respondents' satisfaction with unorthodox monetary policy measures. The authors use ordered probit models to analyze answers given by professional financial market participants to a survey developed by them with Barclays in 2013 finding out that Fed communicates best according to respondents, followed by BoE, ECB and Bank of Japan.

Schonhardt-Bailey (2013) provides in her book a textual analysis of deliberation on monetary policy from 1976 to 2008 tracking the evolution over time of FOMC's deliberating process. Bailey and Schonhardt-Bailey (2008) analyze Fed's FOMC transcripts with text mining. They focus on credibility of central bank and policy effectiveness. The period 1979-1980 - namely the Miller and Volcker eras, respectively – is considered in order to investigate the role of deliberation on shifting the central banks' strategy from an interest rate setting to targeting nonborrowed reserves as a reaction to the high inflation characterizing that years. Once latent topics are identified analyzing transcripts with the software Alceste, the strength of the relation between different topics and speakers is measured according to chi-squared distances, under the assumption that observing distributions of words deviating from a theoretical one in which tokens are conditionally independent provides evidences in favor of an association between individuals themselves and underlying topics. Observing the evolution of the debate over time -that is, the addressed topics and members raising them- the authors point out that in a first stage the chairman Volcker was effective in persuading some of the skeptical FOMC members on the role of money on inflation; then, he moved the discussion to a greater attention on the importance of central bank credibility in pursuing its commitment to maintain low inflation.

Beyond the forward guidance dimension, central bank transparency is considered to be a fundamental principle *per se*. Because of accountability reasons, credibility and transparency

concerns and under the pressure of US institutions and public opinion, since November 1993 the Fed started publishing transcripts of FOMC meetings with a five-year lag and provided the past ones. During the 70's FOMC's members were unaware of their whole statements being recorded and stored during the meetings in order to help the preparation of minutes. According to Hansen et al. (2017) this is the perfect environment for a natural experiment in which they exploit the structural break occurred in November 1993 in order to investigate if a greater awareness by FOMC members about the public dimension of the debate changes the way the discussion[22] evolves and the commission deliberates. Within the literature on transparency, accountability and career concerns (see Holmström, 1999) the authors analyze the verbatim obtained from the transcripts of discussions recorded between 1987 and 2006 (the Greenspan's tenure). On the one hand, transparency is supposed to be a mean through which the principal monitors the agent's effort, increasing the latter's discipline expressed as a greater amount of time spent analyzing the economy before each meeting; on the other hand, a possible drawback of transparency could be an higher conformity due to reputational concerns whenever the member of the committee doesn't know her exact type, namely her level of expertise and the amount of self-knowledge (see Scharfstein and Stein, 1990). To assess the balance between the two effects, a difference in difference regression is estimated including transparency as a dummy variable which takes value 1 after November 1993, while the number of years a member spent at Fed represents the level of experience (since reputation concern is supposed to decrease with experience). An interaction term between the two variables is also considered. In order to estimate the "diff-in-diff" model, qualitative data represented by statements are transformed in quantitative ones, to be included in the model as dependent variables. FOMC's members statements are processed using the Latent Dirichlet Allocation (LDA) in order to identify the topic of each sentence and the relative importance of a specific word. LDA is a clustering algorithm proposed by Blei et al. (2003) assuming each text to be composed by a set of topics to identify (the number of whom is an input set by the researcher); the text is tokenized and divided in documents, where each document is a list of words. Words are grouped into topics and each word can be assigned to many topics. The probability distribution of a topic over the set of words and of each document over topics is provided; doing so, the probability of a given word to appear in the specific topic is obtained for each topic. Moreover, the relative importance of each topic as it is perceived by each speaker is estimated considering the fraction of words related to a specific topic over the total amount of words: that is, the time devoted by a FOMC member to a given topic. The authors estimate the topic models based on the transcripts of the whole meeting, while the statements released in two specific sessions of the meeting, FOMC

22 FOMC members owe to discuss about monetary policy exclusively in formal meetings according to Government in Sunshine Act (1976).

1 and 2, are analyzed to focus on the discussion about economic situation and policy decision. The authors found evidence about the fact that less experienced committee members after transparency tend to gather more information and refer more often to quantitative data (more discipline) but are also more likely to disengage with the discussion.

Hansen and McMahon (2016) apply computational linguistics techniques to statements of the FOMC distinguishing three dimensions of monetary policy communication from the central bank. The first dimension concerns the current monetary policy and the decision rule followed in setting the interest rate and refers to the branch of econometric literature investigated by Stock and Watson (2001) and Bernanke et al. (2005), among the others, and including the interest rate as a key variable affecting the economy consistently with the Taylor's (1993) reaction function. In this extent, the stance of current monetary policy consists with the use of the shadow rate as in Black (1995) and Wu and Xia (2016). The second dimension refers to FOCM own views on the state of economy under the hypothesis that insights are delivered to the market in order to reduce the strength of information asymmetry due to time lags in the availability of data. Given the fact that markets react to central bank announcements (Gürkaynak et al., 2005, Swanson, 2017), the authors estimate the extent in which communication matters and identify the content of the message. FOMC's members statements are processed using the Latent Dirichlet Allocation (LDA) in order to identify the topic of each sentence and the relative importance of a specific word. After identifying the main topics regarding different aspects of the economy (for instance job market, inflation and aggregate demand), few topics of interest are chosen and sentences related to those topics are considered. Using dictionary methods as in Loughran and McDonald (2011), the tone of words contained on sentences related to each topic is assessed to be positive or negative and a time-series balance measure of the FOMC statement on the economic situation is created. The third dimension is about forward guidance. It concerns those statements supposed to be related to future decisions on interest rate according to a narrative approach with manual checking. Words in those Sections are classified as expansionary, neutral or contractionary, the share of the statement dedicated to guidance is quantified and the degree of uncertainty of the statement is measured via dictionary methods (see also Tetlock, 2007 and Schmeling and Wagner, 2019). A FAVAR (Bernanke et al., 2005, Stock and Watson, 2005 and Marcellino et al., 2005) is built up in two steps[23] using the principal components to estimate four factors from a set of 76 variables and including the three dimensions of monetary policy announcements in the vector of driving observed variables. Through the impulse response function (IRF), the effect on the economic variables of a shock hitting one of the three dimensions is investigated (unlike the role of the interest rate is generally emphasized). As far as the financial

23 The factors Ft are estimated using principal components and the VAR in the estimated F_t and the observed Y_t.

market is concerned, the authors conclude that FOMC's statements on the forward guidance dimension referring to future interest rates seem to be more effective than communications on economic conditions, while the effect on real economic variables is weak. Nevertheless, the effect of communication is negligible.

Finally, Acosta (2015) uses LSA and cosine similarities to reduce dimensionality[24] and to compare transcripts and minutes of FOMC meetings in order to measure how transparency evolved over time, being transparency defined as the usefulness of the minutes in interpreting the content of the meeting. The author also investigates how awareness of higher transparency affects the quality of the meeting itself and the informative content of it. It is found out that transparency increased heavily at the end of the 70's mainly because of higher quality of minutes rather than discussion being shaped by members' awareness. On the other hand, consistently with Hansen et al. (2017), conformity increased reducing disagreements within members of the committee.

2.1.3 Text Mining and Central Banking

Apart from BoE, Fed and ECB, which communication strategy has been deeply studied, text mining has been widely applied in analyzing communication for several central banks. As an instance, Bruno (2016) examines the Governor's Concluding Remarks released by Bank of Italy between 1996 and 2015 showing an extremely high level of formality and an index of readability approximately requiring a college degree. The LDA is applied by Shirota et al. (2015) to minutes of monetary policy meetings held at the Bank of Japan (BoJ) between 2012 and 2014. The estimated topics resulting from committee's discussions were consistent with the accommodative policy to which the BoJ committed since 2013 and with the inflation targeting of 2%. Takeda and Keida (2018) discuss results obtained comparing the BoJ's governors' speeches using the LSA model, finding out that speeches are serially correlated themselves while differences among governors persist between different mandates according to the pursued strategy.

2.1.4 ECB's Communication

An early attempt to analyze the semantic content of ECB's statement was conducted by Rosa and Verga (2007) constructing a specific glossary. Between 1999 and 2007 the informative content of introductory statements pronounced during monetary policy presidential press conferences was rather limited, aimed at announcing the adopted policy and sharing the ECB's views on actual and forthcoming economic developments. Rosa and Verga (2007) manually ranked statements according

24 LSA identifies the principal components of the document based on Singular Value Decomposition (SVD).

to the perceived danger for European economy and were able to predict consequent ECB's future monetary policy decisions. Moreover, ECB is able to shape market participants' expectations on short-terms interest rate. An evident limitation of such an approach, as emphasized by the authors themselves, is represented by the broad subjectivity in classifying statements: this is one rationale for applying NLP.

The length of ECB's president statements released during the scheduled press conferences increased over time. As an instance, the length of the policy summary has varied from 58 words on January 2002 to 436 words on December 2015, increasing the attention dedicated to possible future monetary policies (Galardo and Guerrieri, 2017). ECB started providing forward guidance since July 2013 (see and ECB, 2014). Praet (2013) defines forward guidance as "a communication instrument by which central banks convey their monetary policy orientation going forward, conditional on their assessment of the economic outlook". As a matter of fact, central banks influence the economy modifying the short-term interest rate - affecting in this way the long-term structure of interest rates - and providing signals about future policies according to forecasts on the evolution of the economic stance, conditional on the current information set. As the ECB started providing forward, exploiting the potentiality of semantic analysis to investigate the content of ECB's communications and to measure the perception among media and economic agents of the tone of press conferences is a worth exercise. News and media reports contribute to the way agents shape their expectations. ECB's Tobbak et al. (2017) propose two indicators measuring media's perception of ECB's tone at press conferences. The authors compare and index of average hawkishness or dovishness (HD) estimated based on semantic orientation and another index estimated classifying media articles with SVM trained with texts pre-labelled by the them. They also performed an LDA in order to study medias' interpretation and the correlation between debated topics and movements in interest rates in a Taylor rule's environment. It resulted that the index estimated with the SVM was more reliable in detecting the tone of articles as well as more objective. By the LDA the authors proved that medias' coverage moved from monetary policy decisions to unconventional communication on monetary policy. Nevertheless, the estimated HD index results to be highly correlated with MRO and LIBOR rate.

Galardo and Guerrieri (2017) propose an indicator of ECB's verbal guidance on future decisions based on verbal tenses of official statements released in press conferences following monetary policy decisions. Indeed, forward guidance is quantified based on the presence of grammar future markers in the analyzed statements as the ratio between the total number of future tenses and total number of words. Such an approach permits to avoid the use of glossaries reducing the extent of domain dependency and subjectivity. The introduction of forward guidance in 2013 translates in the increasing importance of tokens such as "will" as the Main Refinancing Operations Rate (MRO) gets closer to the effective lower bound. The authors point out that the markets don't merely react to the

announcement of non-standard measures but, rather, expectations on future short terms interest rates captured by variations in the EURIBOR are affected by the way in which the message is conveyed, being those variations negatively related to verbal guidance, used as an instrument to convince agents on the ECB's commitment that the monetary policy will remain accommodative in the near future. The proposed index has the advantage to be continuous over time being based on grammar expressions, while the forward guidance and other unconventional measures appeared recently in the ECB's toolkit.

Similarly to Born et al. (2014), Kahveci and Obadas (2016) perform a semantic analysis of monetary policy statements released by policymakers from Fed, ECB and Central Bank of the Republic of Turkey (CBRT) between 2002 and 2015 comparing the tone and diction of statements using Diction 7 software (working with custom dictionaries created by users) in order to detect the effect of an increasing level of transparency. Like Hansen and McMahon (2016), they consider statements on monetary policy decision from Fed and CBRT as vectors to communicate the chosen policy, the central bank's view about the stance of the economy and to provide forward guidance about future decisions. As a matter of fact, during monetary decision policy statements, Fed and CBRT communicate the policy and the central bank's view about the state of the economy whereas ECB mainly provides the policy decision on interest rate. For that reason, Kahveci and Obadas (2016) perform a semantic analysis of monetary policy statements from the Fed, the ECB and the CBRT and to overcome such heterogeneity the authors use introductory statements released after ECB policy meetings. They focus on two variables, namely certainty and optimism[25], finding out that after the 2007's financial crises the Fed reduced its optimism in statements and increased the level of certainty reflecting forward guidance. The CBRT increased in optimism while ECB remained quite stable in the tone. They also raise the question whether higher optimism can lead to better expectations and faster recover during recession.

As already mentioned, Bholat (2015) emphasized the increasing availability of granular data and the possibility of using them at the BoE. The same argument was stressed by Cœuré (2017) as a member of the executive board of ECB, advocating for the role of central banks as data collectors. On such extent, granular data on individual loans accounting for the 80% of the total balance sheets of European banks are stored through the money market statistical reporting (MMSR). Exploiting such data, ECB plans to provide a new reference rate in order to achieve greater stability due to availability of more reliable benchmarks. AnaCredit project is related to analogous granular data referring to companies rather than individuals. The Payment Services Directive is also aimed at enhancing availability of data on transactions, determining the end of banks' monopoly.

25 The authors define in detail those variables in the appendix of their article.

Big data availability also raises concerns about privacy and confidentiality (see as an instance Stough and McBride, 2014). On this behalf, in April 2016 the EU's (2016) General Data Protection Regulation was approved by the European Parliament. Unlike Fed, ECB doesn't release transcripts of committee meetings. As argued by Hansen et al. (2017), the ECB is "the least transparent of the large central banks" as far as the text availability is concerned, despite its overall reputation is good, however (Rosa and Verga, 2007). The next Section will be devoted to a brief description of ECB's structure with a focus on banking supervision and the availability of public released textual data capable of being studied with text mining techniques will be pointed out. Alternative data sources will be proposed to fill the gap left by the ECB's partial lack of transparency and possibility for future research will be highlighted.

3 ECB's Structures[26] and Texts Availability

The ECB is an EU institution at the heart of the Eurosystem and it is responsible, with the Single Supervisory Mechanism (SSM) and in cooperation with national supervisors, for ensuring that European banking supervision is effective and consistent. ECB formulates the monetary policy for the euro area, that is: ECB takes decisions over interest rates and nonborrowed reserves. The monetary policy decision is explained in detail at a press conference held by the president every six weeks. The President, since 2011 Mr. Mario Draghi, assisted by the Vice-President, chairs the press conference. ECB's main goal is maintaining price stability. To pursue its goals, ECB is organized as follows: the Governing Council (GC) is the main decision-making body of the ECB and usually meets twice a month in Frankfurt am Main (DE). The GC is composed by the six members of the Executive Board and the governors of the 19 national central banks of the euro area. It formulates the monetary policy for the euro area, it defines the general framework under which supervisory decisions are taken, and adopts decisions proposed by the Supervisory Board. It assesses economic and monetary developments and takes its monetary policy decisions every six weeks. To ensure the separation of the ECB's monetary policy and other tasks from its supervisory responsibilities, separate meetings of the Governing Council are held. The Executive Board (EB) consists of the President, the Vice-President and four other members appointed by the European Council. It prepares GC meetings, implements monetary policy for the euro area in accordance with the guidelines specified and the decisions taken by the GC and manages the day-to-day activity of the ECB. It also exercises certain powers delegated to it by the GC, including some of a regulatory nature. Finally, the General Council is composed by the ECB's President and Vice-President and by the governors of national central

26 The main reference for this Section is the "Statute of the European system of central banks and of the European Central Bank", published on the Official Journal of the European Union on 26/10/2012, and the ECB's website: https://www.ecb.europa.eu/ecb/orga/decisions/govc/html/index.en.html

banks. The members of the EB can attend meetings.

Supervision authorities are the European Banking Authority (EBA) and, since 2014, the Single Supervisory Mechanism (SSM). The EBA is an independent EU Authority which works to ensure effective and consistent prudential regulation and supervision across the European banking sector. It is mainly composed by representative of national supervisors. EBA contributes, through the adoption of Binding Technical Standards (BTS) and Guidelines which are adopted by the European Commission, to the creation of the European Single Rulebook in banking. BTS are legally binding and directly applicable in all Member States. The Single Rulebook provides a single set of harmonized prudential rules for financial institutions throughout the EU, helping create a level playing field and providing high protection to depositors, investors and consumers. EBA promotes convergence of supervisory practices and is mandated to assess risks and vulnerabilities in the EU banking sector through regular risk assessment reports and pan-European stress tests. Although the role of EBA is note negligible, from 2014 the main role on supervision is played by SSM. Single Supervisory Mechanism (SSM) refers to the system of banking supervision in Europe, involving ECB and national supervisory authorities, aimed at increasing the resilience of banks. SSM ensures the safety and soundness of the European banking system, pursues financial integration and stability and ensures consistent supervision. Supervision consists in conducting supervisory reviews, on-site inspections and investigations, granting or withdrawing banking licenses, assessing banks' acquisition and ensuring compliance with EU prudential rules. Moreover, SSM can set higher capital requirements ("buffers") in order to counter any financial risks. As far as the decisional process is concerned, the Supervisory Board (SB) prepares the draft decisions and, if the Governing Council does not object within a defined period of time, the decision is deemed adopted. The SB meets twice a month and is composed by the Chair, the Vice Chair chosen from ECB's Executive Board, four ECB's representatives and, finally, the members and representatives of national supervisors.

After the ECB's organization has been briefly depicted, we highlight the main differences between ECB and other important central banks under the point of view of data availability. Starting from 1993 the Fed publishes Minutes six weeks after FOMC meeting and full Transcripts with a 5-years lag. Fed made also available material related to the past, with full transcripts from the 70's and minutes from the 30's. Whereas BoE publishes Minutes but not Transcripts, ECB is considered the least transparent among major central banks, as already mentioned. In details, ECB provides (in pdf version) what follows:

1) Governing Council-Monetary policy decisions from 1999 and to other decisions from 2004:

- "Accounts" of monetary policy decisions meetings from 2015.
- ECB press releases related to monetary policy from 1997.
- Full press conferences from 1998: Introductory statements to monetary policy decisions in

which ECB's provides its views about stance of economy and, since 2013, its expectations.

- Nowadays views on the stance of the economy can also be obtained from accounts of monetary policy meeting, but just from 2015. Full speeches and interviews are available on pdf version from 1997.

2) Committee of Governors (the equivalent of Governing Council at the time of Economic and Monetary Union, before ECB):

- Some authorized record releases between 1962 and 1978.
- Agenda and minutes between 1964 and 1986.
- Further unpublished material can be eventually asked.
- Agenda is also in English while minutes are in French and German.

3) Executive Board (EB):

- Meetings calendar (internal and external).
- Calendar of other meeting the members of EB participate (namely, Governing Council).
- Further textual data available for the Governing Council.

4) Single Supervisory Mechanism (SSM):

- Calendar of meetings of the Chair and Vice-Chair (Danièle Nouy and Sabine Lautenschlager, respectively, between 2014 and 2018).
- Press releases, publications, speeches and interviews. No transcripts or minutes are provided.
- Textual data referring to the member of SSM belonging to EB are available.
- Guide to banking supervision (a general guide).
- Letters to banks: general and specific ones (e.g. 27/11/2017 letter to Dexia Crédit Local about conversion of preferred shares to ordinary shares). Confidential letters sent to banks during the year are not released.

5) European Bank Authority (EBA):

- Meeting Minutes.
- Risk Reports.
- Annual Reports.
- Official speeches.

At the final ring of this hypothetical chain of textual documents, supervised institutions are asked to provide in their financial reports (accounting) information strictly related to banking supervision and financial requirements. As a matter of fact, it emerges that there are three layers of textual data available for research: 1) The decision making of monetary policies, forward guidance and

supervisory decisions (e.g. accounts of meetings and official speeches and interviews), 2) The communication of the decision (e.g. presidential press conferences for monetary policies, letters and guidelines to banks for banking supervision), 3) The interpretation of policies and the implementation of the requirements by supervised entities (as from annual reports). The relevance of what was argued lays in the fact that, if SSM requests are not fulfilled, financial institutions are punished: punishments are fines, usually taking the form of capital or liquidity requirements. In such view, the possibility of applying ML tools to define a stable semantic and quantitative relation among the documents toward the three layers is challenging.

In the next section, we investigate in which extent it is possible to address future research at predicting the content (or at least the direction) of the supervisory guidelines and with which time-lag. Moreover, the question about the way in which the composition of the committees affects the final decision spontaneously arises. We will implicitly assume that policy makers in the discussion phase represent their origin countries[27]. National preferences can be replaced, in the lack of transcripts and minutes of the committee meeting (the first ring of the chain), by the ones expressed by a set of national influencers (national banks governors, national supervisors, members of ECB and SSM) via speeches or social media. Our approach is aimed at uncovering "hidden truths" (Meade and Acosta, 2015) as we believe that supervisors can send messages to agents and strengthen cooperation between different authorities (Born et al., 2014 and Oosterloo et al., 2007). Moreover, we believe that NLP can reveal policymakers' preferences (Tobbak et al., 2017): in fact, final guidelines issued by the SSM are mixture of influences coming from different entities and in the absence of meeting minutes it is almost impossible to identify individual contributions. Instead, interviews and speeches are more flexible than official documents, as their schedule is less rigid, are more likely to contain surprising elements (Born et al., 2014 and Rosa and Verga, 2007) and allow to assess, at least approximately, any single country contribution to the final decision. Future research can be devoted to the creation of a richer database of influencers' public statements proxying their contribution to the decision process.

4 Describing SSM's agenda with topic modelling

We created a corpus of official statements released by European influencers by scraping ECB's and SSM's official websites with Beautiful Soup[28] and Python. The whole list of potential influencers,

27 SSM's main components are actually from Belgium, Finland, France, Germany and Italy.

28 https://www.crummy.com/software/BeautifulSoup/bs4/doc/

their positions and appointment periods[29] are reported in Table A1. Documents were aggregated by country and by period, so that each country represented a single corpus for each year between 2014 and 2017. At first, documents were tokenized and pre-processed: a domain-specific list of stopwords was created by the authors. As an instance, because we are investigating supervisory decisions, and the guidance on non-performing loans has been at the core of the debate in 2016 and 2017, the single token "non" shouldn't be included in the set of stopwords as, instead, would be in other analysis. Similarly, considering that we are dealing with documents concerning financial and monetary topics, terms like "up" or "down" can't be neglected too. Moreover, a specific dictionary containing typical financial abbreviations and acronyms ("NPLs", "SSM", "ECB" and "ROA", for instance) was generated integrating the one implemented in qdapDictionaries[30], to end up with a set of 123,078 words. Those terms were compared with the tokens obtained from our corpus in order to remove errors and proper names. All the texts were in English[31] and were stemmed after removing stopwords; the resulting corpus were converted to document-term matrices (DTM) of bigrams. The choice of tokenizing using bigrams rather than more common unigrams was due to the importance of compound terms in financial and macroeconomic domains[32]. The obtained DTM was used to feed a 5-topics[33] LDA for each country and for each year. The LDA was estimated using a collapsed Gibbs sampling algorithm with 4000 iterations.

Probability distributions over most frequent bigrams for each estimated topic are reported in the appendix, for each year, for Belgium, Finland, France, Germany, Italy and Portugal. It should be noticed that the first five countries have always had a representative in the SSM either as board members, president or vice-president, while the ECB vice-president between 2010 and 2017 has been the Portuguese Victor Constancio. Since no enough texts were available for Finland between 2014 and 2016, further public statements from individual influencers were scraped from the English edition

29 We write "NO INFO" when the information about the appointment period was not available.

30 https://cran.r-project.org/web/packages/qdapDictionaries/index.html

31 The only exception was the interview released by Ignazio Angeloni with "Il Sole 24 ore" on 26/06/2018 and available just in Italian. In order to be as objective as possible, we translated it automatically with the online translator DeepL, which is based on deep learning algorithms.

32 Think to bigrams such as "asset quality", "income statement", "balance sheet", "non-performing", "quantitative easing" or "big data". It should also be noticed that the majority of those expressions contain words that, if read singularly, have negligible informative power.

33 The choice of the number of topics is rather debated in the literature, although there is not an established practice. To some extent, it depends on the problem at hand. As an example, if we are concerned with estimating the number of topics in a corpus obtained from an encyclopedia (which likely has a high variability as there are thousands of covered topics) the parameter for the number of topics k should be fairly high. On the contrary, dealing with documents referred to relatively close topics (such as financial stability, banking supervision, financial risk and so on) if k is too high, the estimated topics will be quite similar among themselves: for such a reason, we choose a relatively small k. Nevertheless, a more rigorous heuristic approach to choose the number of topics was proposed by Zhao et al. (2015).

of Reuters web site and just articles containing the name of the influencer in the title were considered. In this preliminary work we mainly present results referring exclusively to texts available at official ECB and SSM web sites. We plan, for future applications, to enlarge the set of potential influencers.

LDA provides useful hints to describe the political debate over time, although we applied it statically for single years, single countries and fixed topics. We selected the most relevant estimated topics in 2014 and in 2017 in order to highlight how issues addressed by policymakers' evolved. Figures 1-10 report, for illustrative purposes, the word clouds in which the size of terms represents their probability to appear in the estimated topics.

Our main finding is that bigrams linked to NPLs appear for the first time in 2014, in a topic estimated over the French corpus (Figure 1). NPLs were not at the core of the debate in 2014, as they don't appear among relevant tokens, for any other country, and still in 2015 they appeared only in the estimated topics 3 and 4 of France and Belgium, respectively (see the appendix). According to the estimated LDA, NPLs became a common "hot" topic just since 2016 (in the appendix: topic 4 of Italy, topic 3 of France, topic 1 of Germany and topic 5 of Belgium). Looking at official documents, the expression "non-performing loans" appeared just for three times in the late March 2015 in the ECB's (2015) 83-pages annual report of on supervisory activity, while NPLs were indicated as a supervisory priority by ECB (2016) just in December 2016. A public consultation was launched in September 2016 on the draft guidance to banks on NPLs to end up with the first official guidance in ECB (2017a). Generally speaking, the topic in Figure 1 estimated based on official speeches and interviews released by French influencers, concerns the asset quality of European supervised entities and their exposure to NPLs. The term "tier", as an instance, regards the Basel III prescriptions about the composition of capital and on the percentage of equity and reserves on risk weighted assets (BIS, 2010). Interestingly, Benoît Cœuré (2013), a French member of the Executive Board of the ECB, in October 2013 argued that:

> "It seems therefore useful to look also at East Asia's response to its crisis and see what lessons can be learned. […] Non-viable financial institutions were closed, viable institutions recapitalised and strengthened, value-impaired assets dealt with and the corporate sector restructured, including through foreign investment. These measures […] were associated with a painful adjustment process in the crisis countries but, at the same time, they made a swift recovery possible. This experience contrasts with Japan's experience in the late 1980s. Following the bursting of Japan's asset bubble, there was by and large no rapid write-down of non-performing assets. Due to the low profitability of banks, it was believed that an immediate write-off of bad loans would prevent compliance with capital requirements. The recognition of losses was postponed and, as a

consequence, capital continued to be allocated to investments with limited positive impact on Japan's long-term growth potential, which some have called "zombie lending".

Such a statement already contains, in the subtext, the seed of the forthcoming measures aimed at forcing accounting provisions and write-off practices with reference to NPLs (especially unsecured and partially unsecured exposures, according to their vintage category) in order to strengthen banks' balance sheets (see ECB, 2017a, b). Those findings provide qualitative and quantitative evidence in favour of France placing ECB's attention on banks' NPLs. Noteworthy, as all the texts that we used to feed the LDA were publicly available in real-time, there was room for supervised banks to anticipate, at least barely, the direction of supervisory agenda; nonetheless, if a stable relation between specific countries' political agenda and the final policy is established, variations in the composition of the authority's boards or shifts in the measures advocated by countries or groups can be predictive of shifts in supervisory practices.

The topic in Figure 2 is estimated based on official statements from Belgian influencers; it indicates the concerns for low inflation, and the need for expansive monetary policies. Below we report an extract from Praet (2014):

> "Inflation developments continue to surprise on the downside and recent weakness in wage growth casts doubt on the expected strengthening of domestic price pressures. Inflation in the euro area has declined from a high of more than 3% at the end of 2011 to a low of 0.3% in September. And the volatility we have seen recently in medium- to longer-term inflation expectations is a cause for extra vigilance."

Low and eventually negative inflation between 2013 and 2014 due to weak aggregate demand and falling oil prices was at the center of the debate, with concerns for possible de-anchoring of inflation expectations and policymakers and the academics advocating for government bonds and corporate assets purchases[34] as a possible solution (see Conti et al., 2015).

34 See https://voxeu.org/article/low-inflation-eurozone

Table A1: List of influencers by country			
Countries	Individuals	Position/Institution	Appointment period
Belgium	Luc Coene	SSM	2015-2016
Belgium	Tom Dechaene	SSM and Central bank of Belgium	NO INFO
Belgium	Peter Praet	ECB, Executive board	Since 2011
Belgium	Jan Smets	Central bank of Belgium, governor	Since 2015
Finland	Pentti Hakkarainen	SSM	Since 2017
Finland	Olli Rehn	Bank of Finland, governor	Since 2016
Finland	Anneli Tuominen	SSM	Since 2015
Finland	Mervi Toivanen	SSM	NO INFO
Finland	Sirkka Hämäläinen	SSM	2014-2016
France	Danièle Nouy	SSM + Steering Committee	Since 2014
France	François Villeroy de Galhau	Central bank of France, governor	Since 2015
France	Denis Beau	SSM	Since 2014
France	Benoît Cœuré	ECB, Executive board	Since 2012
Germany	Sabine Lautenschläger	SSM + Steering Committee	Since 2014
Germany	Jens Weidmann	Deutsche Bundesbank, president	Since 2011
Germany	Joachim Wuermeling	Deutsche Bundesbank, board	Since 2016
Germany	Felix Hufeld	SSM + Steering Committee	Since 2015
Italy	Ignazio Angeloni	SSM	Since 2014
Italy	Fabio Panetta	SSM + Steering Committee	Since 2014
Italy	Mario Draghi	ECB, president	Since 2011
Italy	Ignazio Visco	Bank of Italy, governor	Since 2011
Portugal	Vítor Constâncio	ECB, Vice-president	2010-2017
Portugal	Carlos Costa	Bank of Portugal, governor	Since 2010
Portugal	Elisa Ferreira	SSM	Since 2016
Spain	Luis de Guindos	ECB, Vice-president	Since 2018
Spain	Margarita Delgado	SSM	Since 2014
Spain	Pablo Hernández de Cos	Central bank of Spain, governor	Since 2018
Others			
Canada	Julie Dickson	SSM	2014-2016
Ireland	Philip Lane	Central bank of Irelnd, governor	Since 2015
Ireland	Ed Sibley	SSM + Steering Committee	Since 2017
Greece	Yannis Stournaras	Central bank of Greece, governor	Since 2014
Greece	Ilias Plaskovitis	SSM + Steering Committee	NO INFO
Netherlands	Frank Elderson	SSM	Since 2018
Netherlands	Klaas Knot	Dutch central bank, president	Since 2011
Austria	Andreas Ittner	SSM	Since 2014
Austria	Helmut Ettl	SSM	Since 2014
Austria	Ewald Nowotny	National bank of Austria	Since 2008
Cyprus	Yiangos Demetriou	SSM	Since 2014
Cyprus	Chrystalla Georghadji	Central bank of Cyprus, governor	Since 2014
Slovak Republic	Jozef Makuch	Central bank of Slovak Republic, governor	Since 2010
Estonia	Madis Müller	SSM	Since 2014
Estonia	Kilvar Kessler	SSM	Since 2014
Estonia	Ardo Hansson	Bank of Estonia, governor	Since 2012
Luxembourg	Norbert Goffinet	SSM	NO INFO
Luxembourg	Claude Simon	SSM	NO INFO
Luxembourg	Yves Mersch	ECB, Executive board	Since 2012
Luxembourg	Gaston Reinesch	Central bank of Luxembourg, governor	Since 2013
Latvia	Pēters Putniņš	SSM	NO INFO
Latvia	Ilmārs Rimšēvičs	Central bank of Latvia, governor	Since 2001
Latvia	Zoja Razmusa	SSM	Since 2014
Lithuania	Vytautas Valvonis	SSM	NO INFO
Lithuania	Vitas Vasiliauskas	Central bank of Lithuania, chairman	Since 2011
Malta	Oliver Bonello	SSM	NO INFO
Malta	Catherine Galea	SSM + Steering Committee	NO INFO
Malta	Mario Vella	Central bank of Malta	Since 2016
Slovenja	Primož Dolenc	SSM	NO INFO
Slovak Republic	Vladimir Dvořáček	SSM	Since 2014

Notes: List of influencers, countries and positions.

Figure 1: Word cloud for France, 2014

Figure 2: Word cloud for Belgium, 2014

Figure 3: Word cloud for Germany, 2014

Figure 4: Word cloud for Italy, 2014

Figure 5: Word cloud for Portugal, 2014

Figure 6: Word cloud for France, 2017

Figure 7: Word cloud for Belgium, 2017

Figure 8: Word cloud for Germany, 2017

Figure 9: Word cloud for Italy, 2017

Figure 10: Word cloud for Portugal, 2017

Figure 3 reflects German emphasis on regulatory stuffs and on the framework of supervisory activities: see as an instance the recurrent presence of tokens referring to "level playing field", reminding to the idea of a fair competition in the banking sector enhancing stability and growth. In fact, as stated by Sabine Lautenschläger (2014):

> "European legislators and the public, too, expect the SSM to ensure a resistant, robust banking system in the euro area. Supervision in Europe is rightly expected to be better than what 19 national supervisors could achieve with national means. The new system of supervision should be neutral and not wedded to national thinking and traditions. It should, where appropriate, produce a level playing field. It had become apparent in the financial crisis that the difficulties of large, strongly interlinked banks were not just a national problem and that difficulties can spill over into the public sector and the real economy."

Moreover, European banking supervision is rooted on the activity of a single resolution authority aimed at minimizing possible systemic externalities of single banks restructuring.

According to Figure 4 and 5, Italian and Portuguese policymakers were in 2014 particularly concerned about the stance of the economy and the efficacy of monetary policies when the central bank is constrained at the zero-lower bound. Effectiveness of unconventional monetary policies -that is, the ability to model inflation expectations- and a sound prudential supervision are also closely related in stabilizing "growth" in the "euro area" and in reducing the extent of imperfect information, which undermine the rational expectations assumption (see Moessner et al., 2017, Swanson, 2017 and Wu and Xia, 2016).

Finally, considering the topics describing the debate in 2017, the scenario is less heterogeneous. As shown in Figure 6-10, tokens related to NPLs and credit risk appear in the estimated agenda of all the most influential countries, meaning that from 2014 the debate started to converge to a shared "hot" topic. Interestingly, the token "hard brexit" enters in the German cloud and appears to be related to "npls" and "financial stability" as a potential source of uncertainty and, consequently, financial risk (Figure 8). Consistently, the Belgian word cloud in Figure 7 refers to concerns for international trade, risks of deflation, potential uncertainty and, to finalize, financial instability.

5 Conclusions

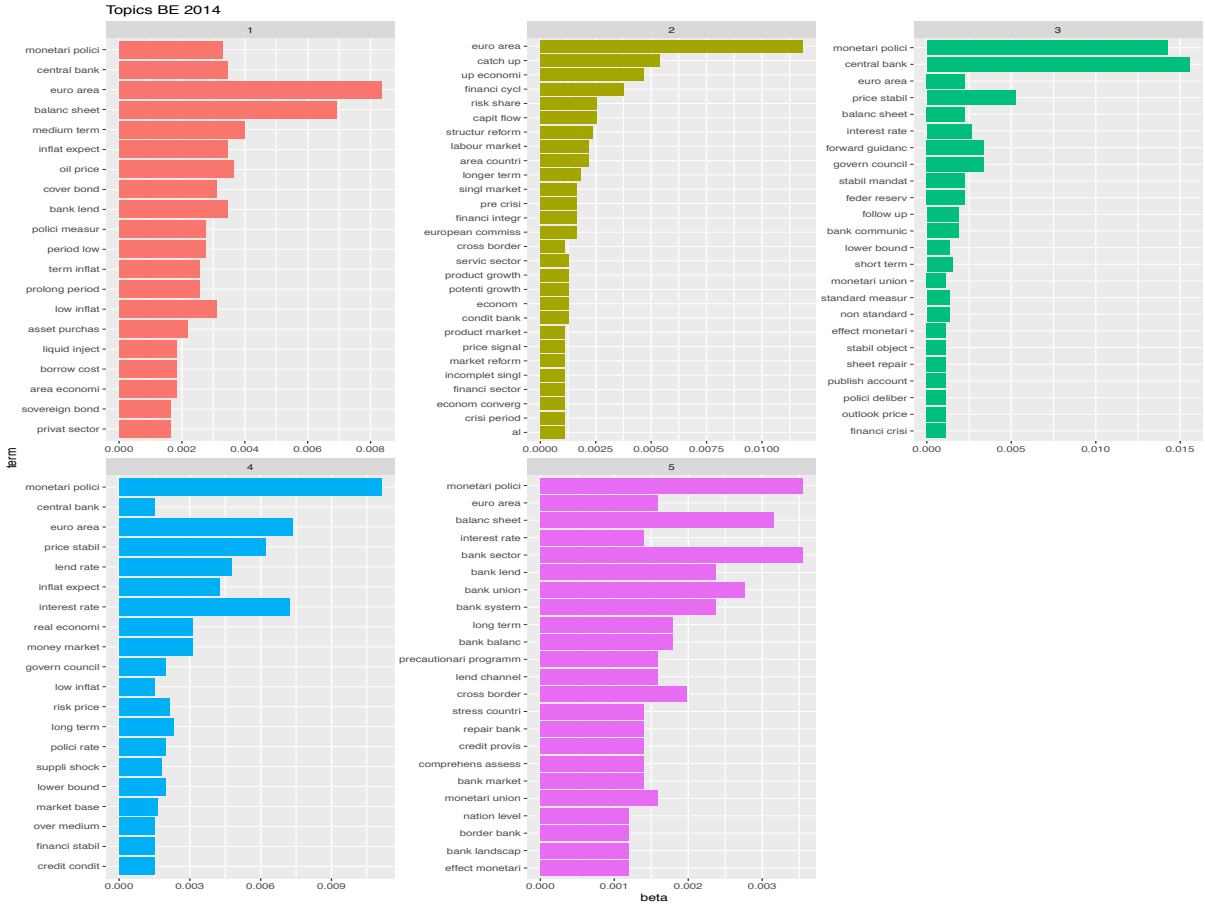
Communication strategies adopted by the most influencing central bankers are evolving in the last decade, and the ECB is no exception. Unconventional monetary policies pursuing financial, economic and inflationary goals are based on communication strategies aimed at modelling expectations and reducing the degree of uncertainty due to information asymmetries of agents. The

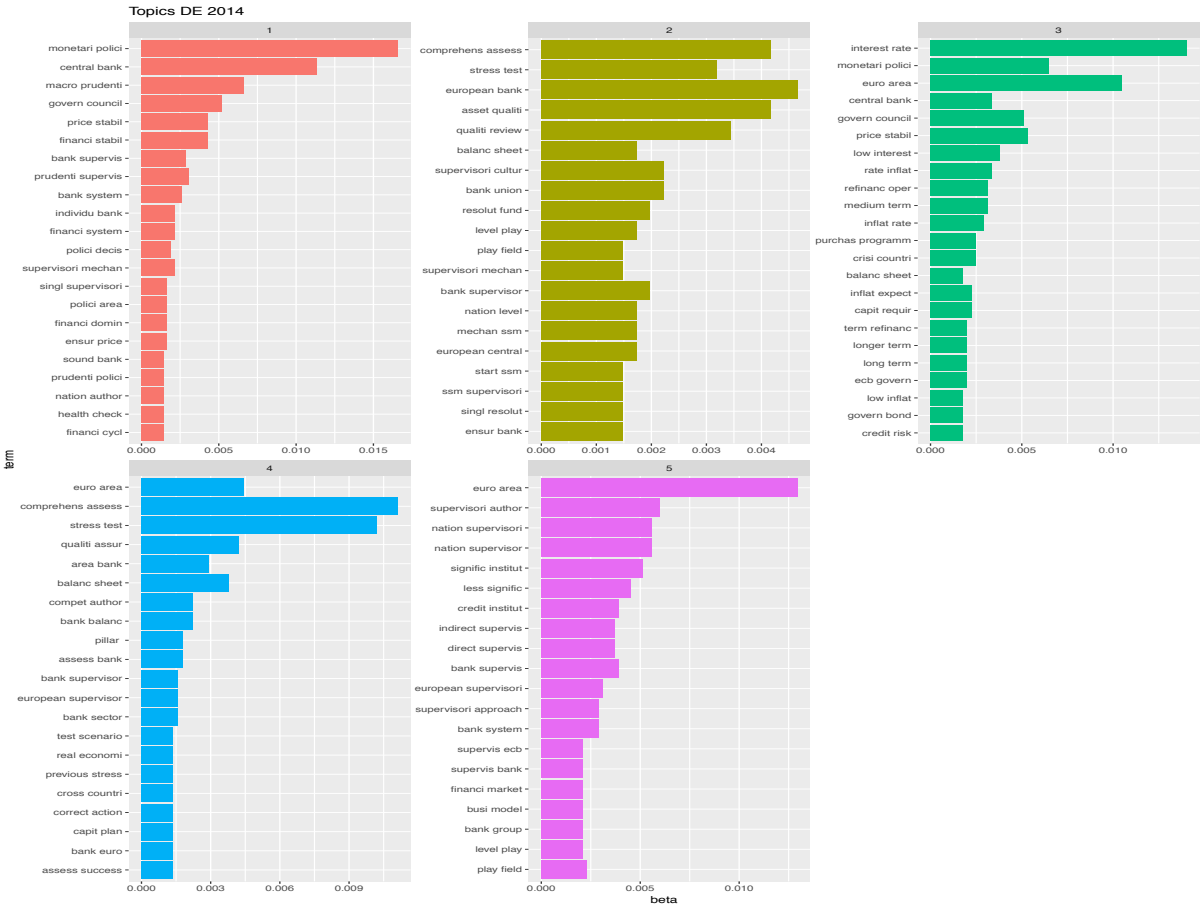
increasing complexity of economic phenomena, the high dimensionality of data and the need to combine textual and numeric information, make traditional approaches inadequate. Machine learning and text mining are proving to be useful tools in analyzing big data and retrieving quantitative insights from texts. Text mining is extremely powerful in investigating how institutions shape their strategies and the extent in which public debate and internal discussions affect the resulting policies and the following market's reactions. Consequently, there is potential for dramatic gains by exploiting text mining to address communication and decision making-related topics, especially for central banks. In the present work we took stock of what has been done so far in the literature exploiting text mining to analyze central banks' communication and we provided instances of the way central banks themselves use machine learning and text mining. Furthermore, having established that central bankers are increasingly adopting communication strategies as unconventional monetary tools -ie, sending signals through forward guidance- we have presented the main research strategies adopted by researchers and professionals to quantify, via text mining, the effects of those practices. Furthermore, we have proposed to apply topic modelling, and especially the LDA, to describe qualitatively and quantitatively the political agenda within the European authority for banking supervision and to disentangle the contribution of any country to the final regulatory outcome. National preferences can be replaced, in the lack of transcripts and minutes of the committee meeting, by the preferences implicitly expressed by policymakers via speeches or social media. Our approach is aimed at uncovering "hidden truths", as we believe that supervisors can use communication to send messages to supervised entities and to different supervisory authorities (Born et al., 2014 and Oosterloo et al., 2007). We have shown how NLP can reveal policymakers' preferences through interviews and speeches, that are more flexible than official documents and more likely to contain surprising elements. Moreover, individual speeches and interviews allow to assess, at least approximately, any single country contribution to SSM's guidances to banks.

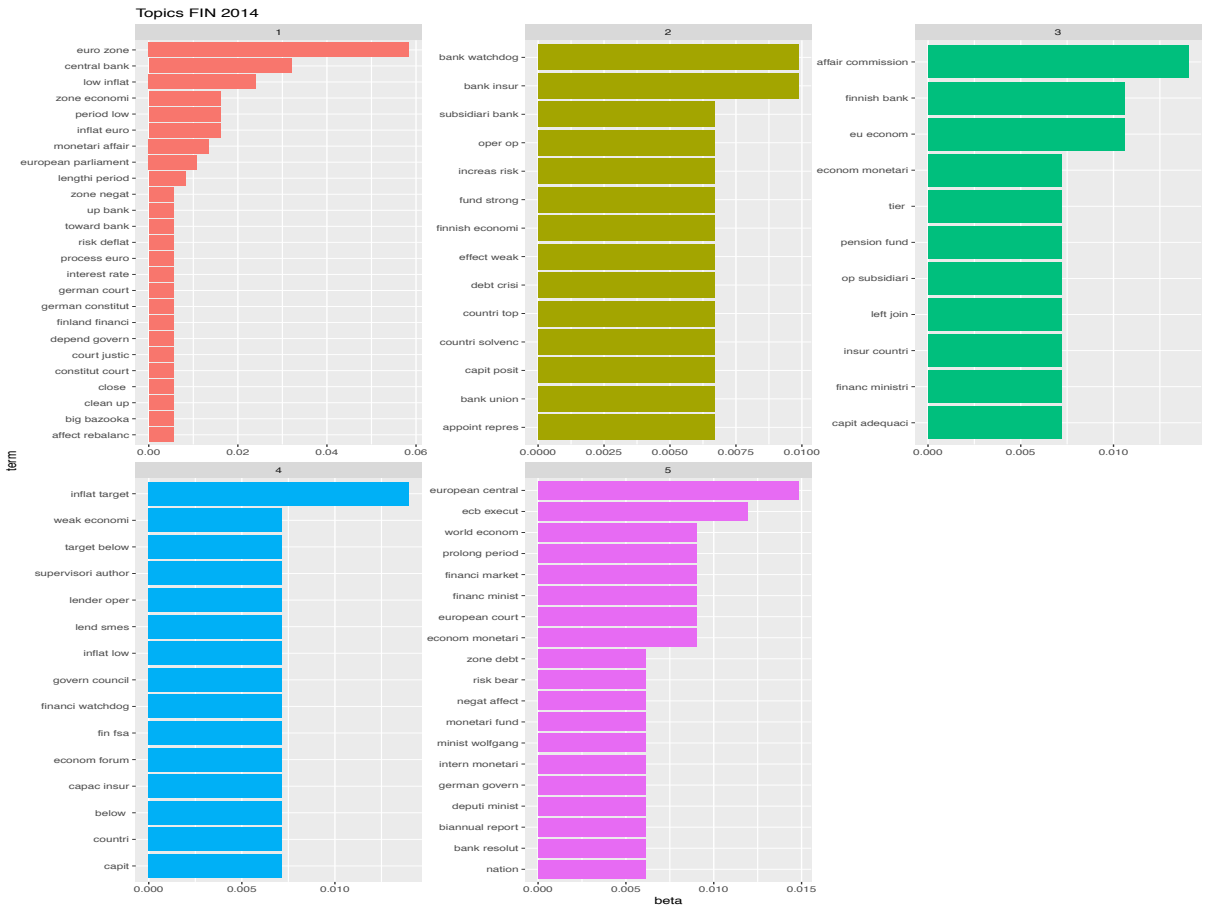
Applying LDA we were able to anticipate the emergence of a crucial topic -that is, NPLs- two years before its inclusion among supervisory priorities. A preliminary result of our work is the evidence about French influencers pushing the debate toward supervisory practices aimed at preventing risks to the financial stability due to NPLs: in fact, already in 2014 (the year in which the SSM started its activity) France was the only country for which tokens referring to non performing exposures appeared with a high probability in an estimated topic. In 2015 NPLs were still at the core of the debate just for France and Belgium and became a shared topic just between 2016 and 2017, in a landscape of measures aimed at improving supervised entities' asset quality. We consider such an evidence the starting point for further research shedding light on the dynamic evolution of the political debate through the use of topic modelling under the assumption that, in banking supervision as well as in monetary policy, policymakers send signals to stakeholders on future actions. In such extent, a

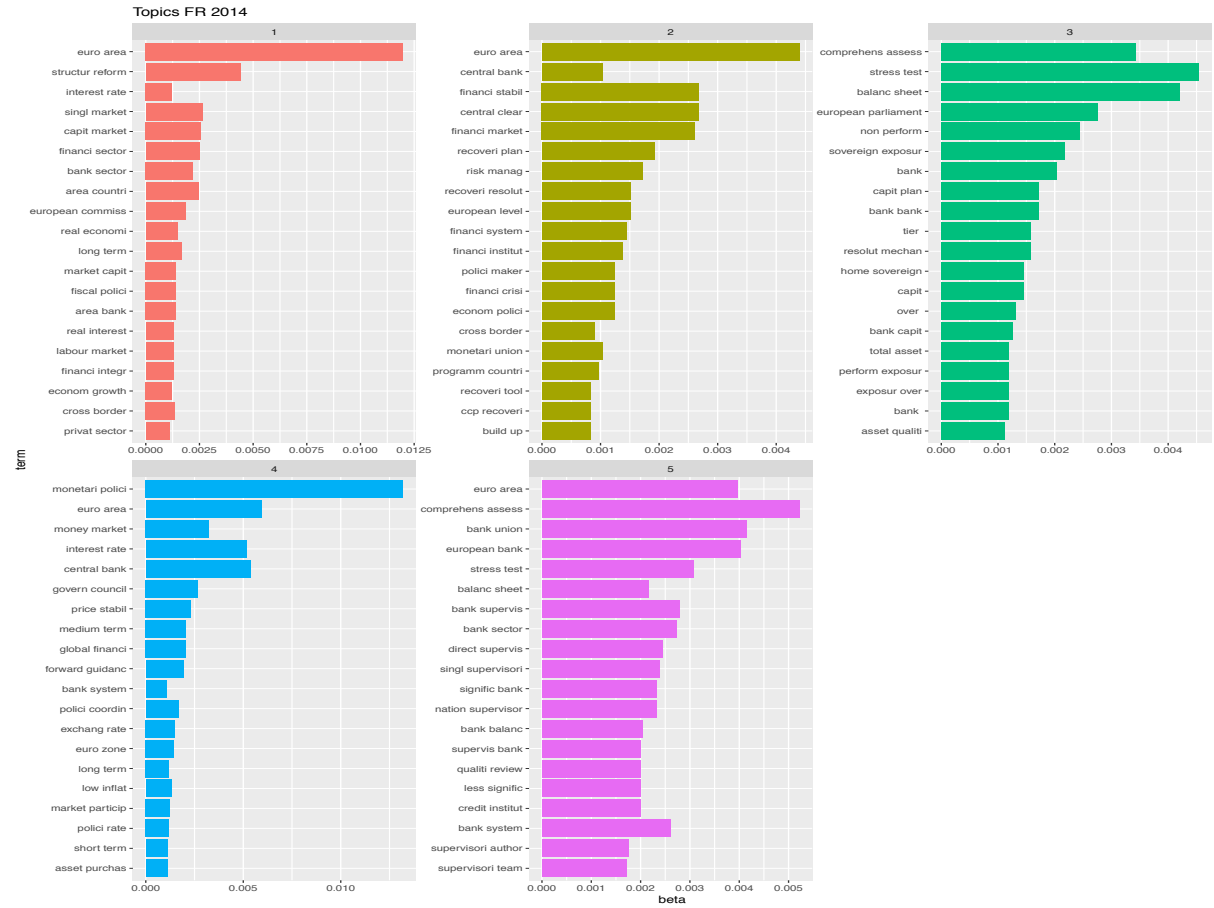
possible strategy for a future in-depth analysis could be applying the dynamic version of the LDA proposed by Blei et al. (2006). Furthermore, we believe supervisory addresses constituting the legal framework of banking supervision to be the result of a "mixture" of influences arising from different countries leading ECB and SSM; consistently, we consider LDA the natural tool to investigate those questions as in that model documents are assumed to be a "mixture" of topics and topics are composed by a "mixture" of terms, having the "mixture" a statistical meaning. Finally, we believe that text mining is a powerful tool allowing supervised entities to identify and anticipate policymakers' preferences in order to fulfill timely supervisory requirements and, for the researchers, to assess the effect of those policies on economic and financial outcomes. We leave to future research the analysis of supervisory requirements and their comparison with the perception by supervised institutions. As a matter of fact, such an approach can be applied to any political decision in which the final outcome is likely to be a mixture of different influences and in which the composition of the deciding authority matters, so that individual contributions need to be assessed.

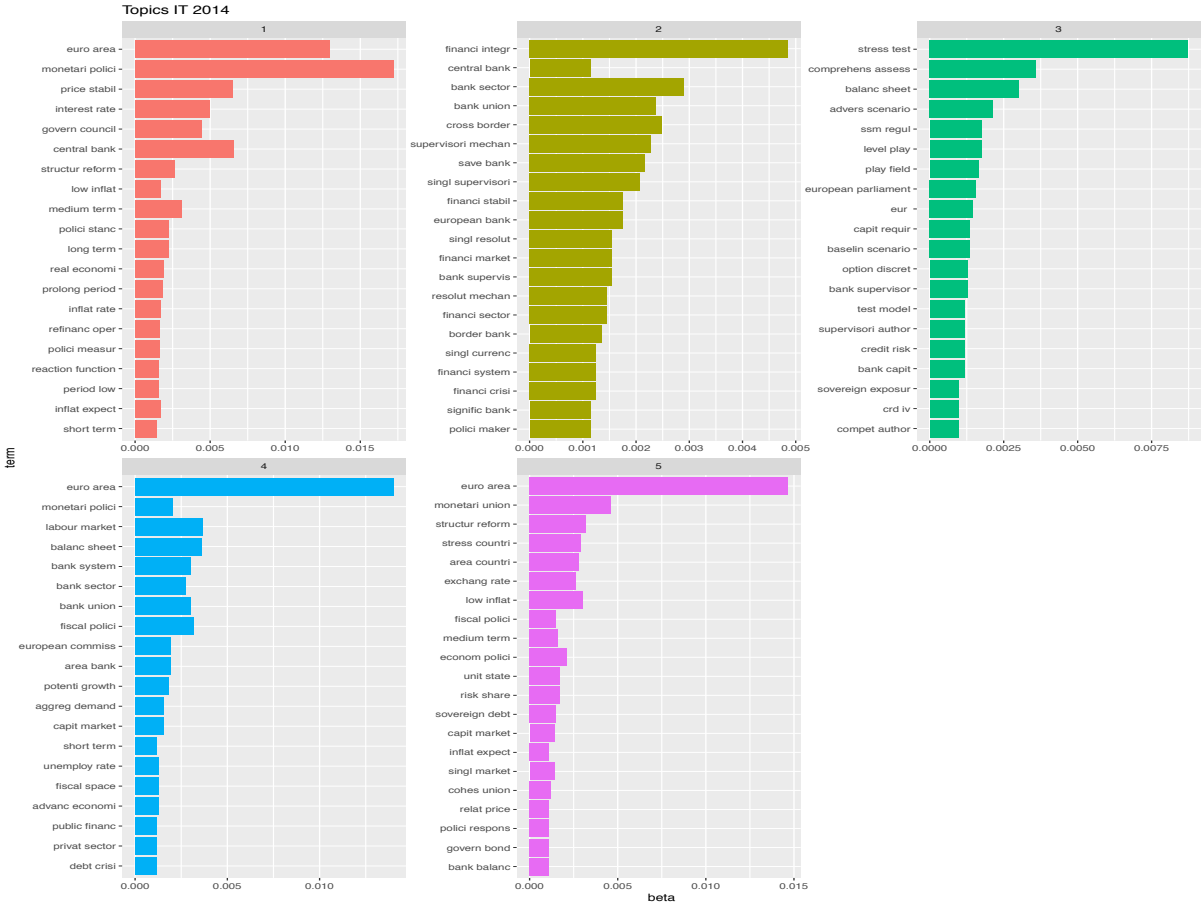
Appendix: estimated "beta" probabilities of terms in topics, 2014-2018

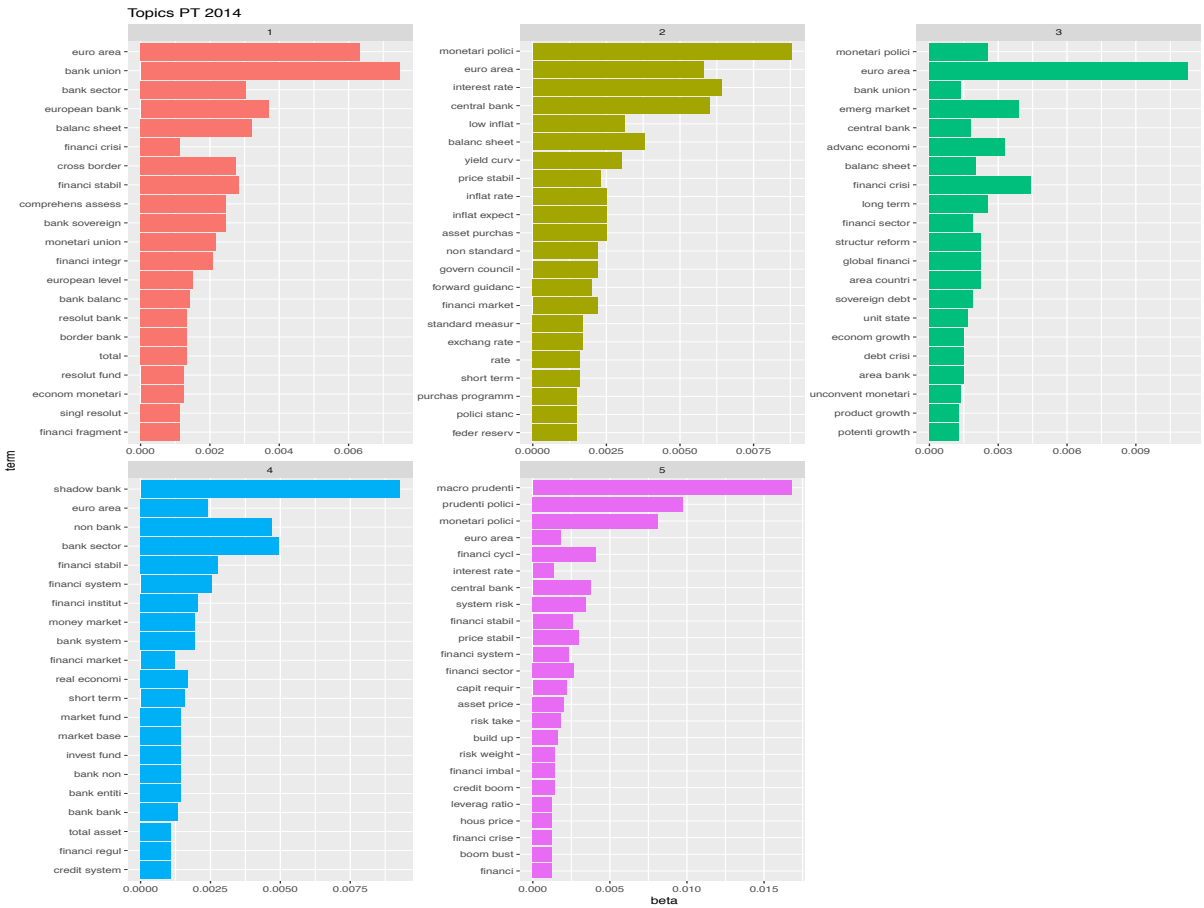


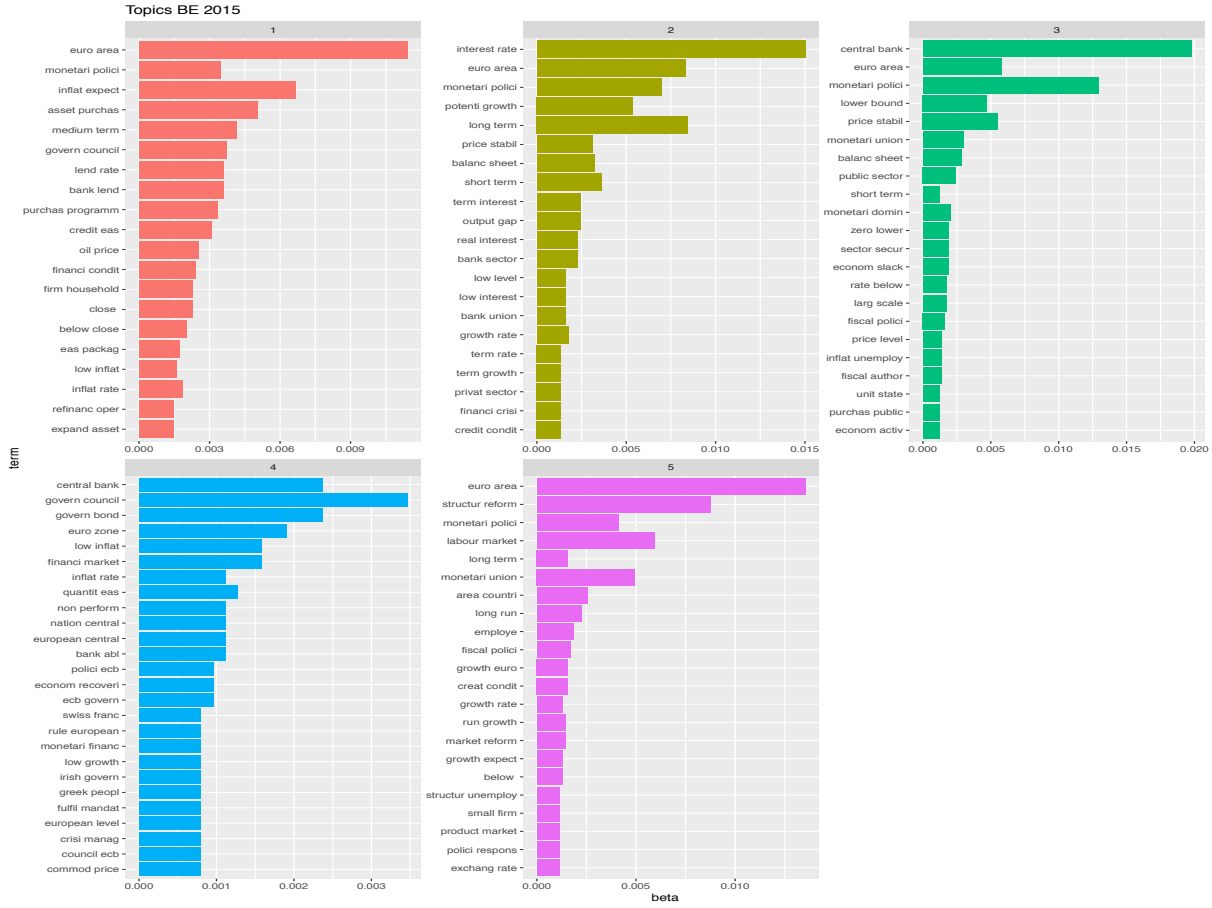


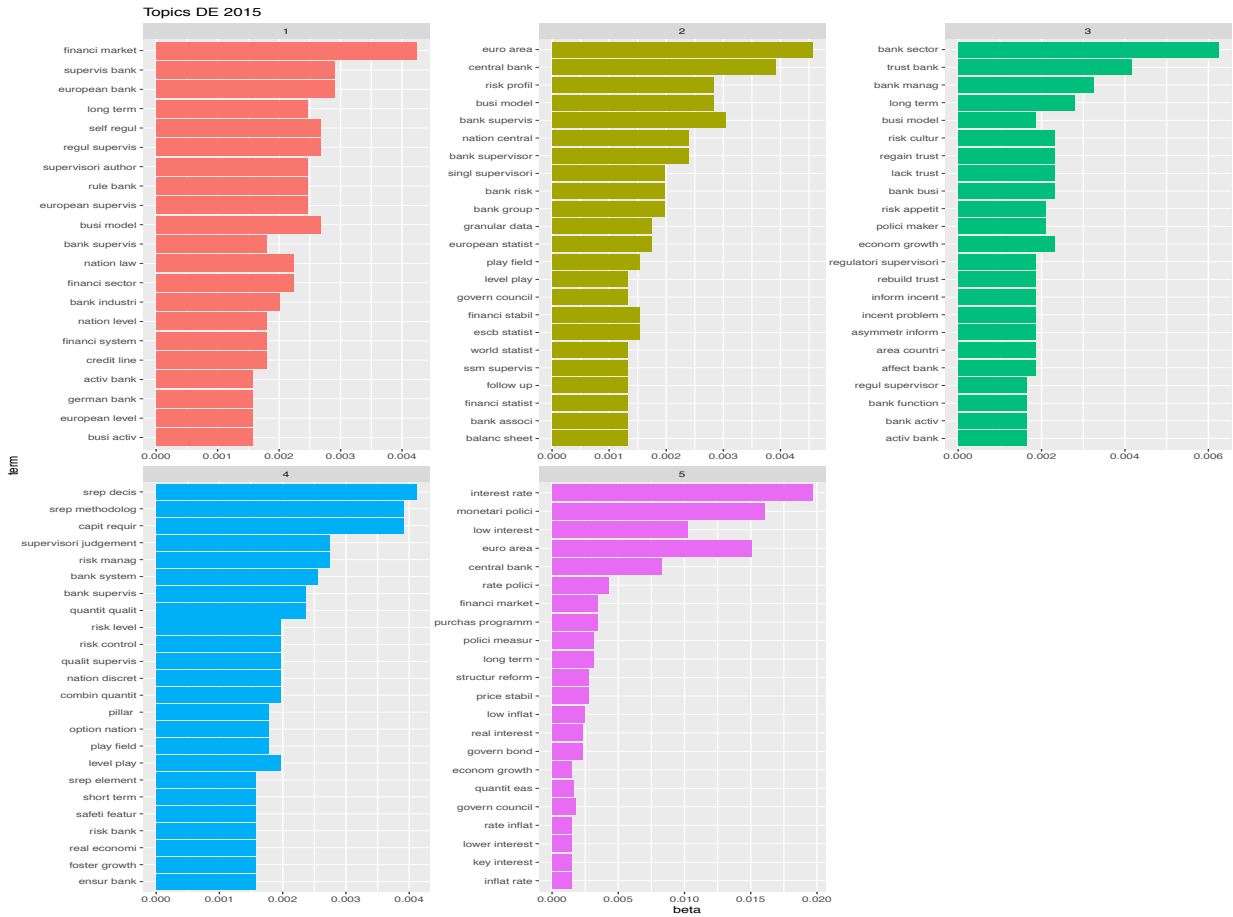

Topics FIN 2014
1
euro zone
central bank
low inflat
zone economi
period low
inflat euro
monetari affair
european parliament
lengthi period
zone negat
up bank
toward bank
risk deflat
process euro
interest rate
german court
german constitut
finland financi
depend govern
court justic
constitut court
close
clean up
big bazooka
affect rebalanc
0.00
0.02
0.04
0.06
2
bank watchdog
bank insur
subsidiari bank
oper op
increas risk
fund strong
finnish economi
effect weak
debt crisi
countri top
countri solvenc
capit posit
bank union
appoint repres
0.0000
0.0025
0.0050
0.0075
0.0100
3
affair commission
finnish bank
eu econom
econom monetari
tier
pension fund
op subsidiari
left join
insur countri
financ ministri
capit adequaci
0.000
0.005
0.010
4
inflat target
weak economi
target below
supervisori author
lender oper
lend smes
inflat low
govern council
financi watchdog
fin fsa
econom forum
capac insur
below
countri
capit
0.000
0.005
0.010
5
european central
ecb execut
world econom
prolong period
financi market
financ minist
european court
econom monetari
zone debt
risk bear
negat affect
monetari fund
minist wolfgang
intern monetari
german govern
deputi minist
biannual report
bank resolut
nation
0.000
0.005
0.010
0.015
beta
term


Topics FR 2014
1
euro area
structur reform
interest rate
singl market
capit market
financi sector
bank sector
area countri
european commiss
real economi
long term
market capit
fiscal polici
area bank
real interest
labour market
financi integr
econom growth
cross border
privat sector
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
2
euro area
central bank
financi stabil
central clear
financi market
recoveri plan
risk manag
recoveri resolut
european level
financi system
financi institut
polici maker
financi crisi
econom polici
cross border
monetari union
programm countri
recoveri tool
ccp recoveri
build up
0.000
0.001
0.002
0.003
0.004
3
comprehens assess
stress test
balanc sheet
european parliament
non perform
sovereign exposur
bank
capit plan
bank bank
tier
resolut mechan
home sovereign
capit
over
bank capit
total asset
perform exposur
exposur over
bank
asset qualiti
0.000
0.001
0.002
0.003
0.004
4
monetari polici
euro area
money market
interest rate
central bank
govern council
price stabil
medium term
global financi
forward guidanc
bank system
polici coordin
exchang rate
euro zone
long term
low inflat
market particip
polici rate
short term
asset purchas
0.000
0.005
0.010
5
euro area
comprehens assess
bank union
european bank
stress test
balanc sheet
bank supervis
bank sector
direct supervis
singl supervisori
signific bank
nation supervisor
bank balanc
supervis bank
qualiti review
less signific
credit institut
bank system
supervisori author
supervisori team
0.000
0.001
0.002
0.003
0.004
0.005
beta
term

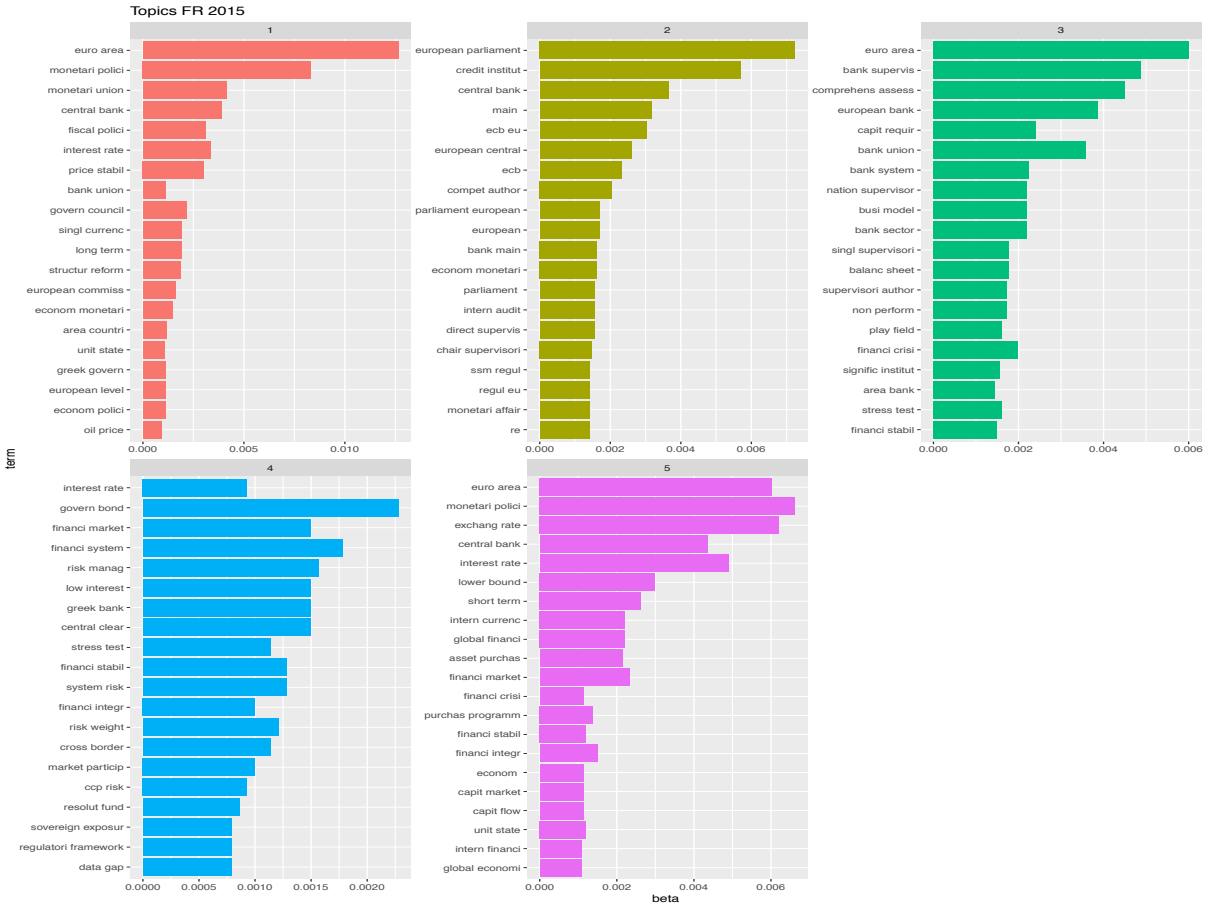

Topics IT 2014
1
euro area
monetari polici
price stabil
interest rate
govern council
central bank
structur reform
low inflat
medium term
polici stanc
long term
real economi
prolong period
inflat rate
refinanc oper
polici measur
reaction function
period low
inflat expect
short term
0.000
0.005
0.010
0.015
2
financi integr
central bank
bank sector
bank union
cross border
supervisori mechan
save bank
singl supervisori
financi stabil
european bank
singl resolut
financi market
bank supervis
resolut mechan
financi sector
border bank
singl currenc
financi system
financi crisi
signific bank
polici maker
0.000
0.001
0.002
0.003
0.004
0.005
3
stress test
comprehens assess
balanc sheet
advers scenario
ssm regul
level play
play field
european parliament
eur
capit requir
baselin scenario
option discret
bank supervisor
test model
supervisori author
credit risk
bank capit
sovereign exposur
crd iv
compet author
0.0000
0.0025
0.0050
0.0075
4
euro area
monetari polici
labour market
balanc sheet
bank system
bank sector
bank union
fiscal polici
european commiss
area bank
potenti growth
aggreg demand
capit market
short term
unemploy rate
fiscal space
advanc economi
public financ
privat sector
debt crisi
0.000
0.005
0.010
5
euro area
monetari union
structur reform
stress countri
area countri
exchang rate
low inflat
fiscal polici
medium term
econom polici
unit state
risk share
sovereign debt
capit market
inflat expect
singl market
cohes union
relat price
polici respons
govern bond
bank balanc
0.000
0.005
0.010
0.015
term
beta

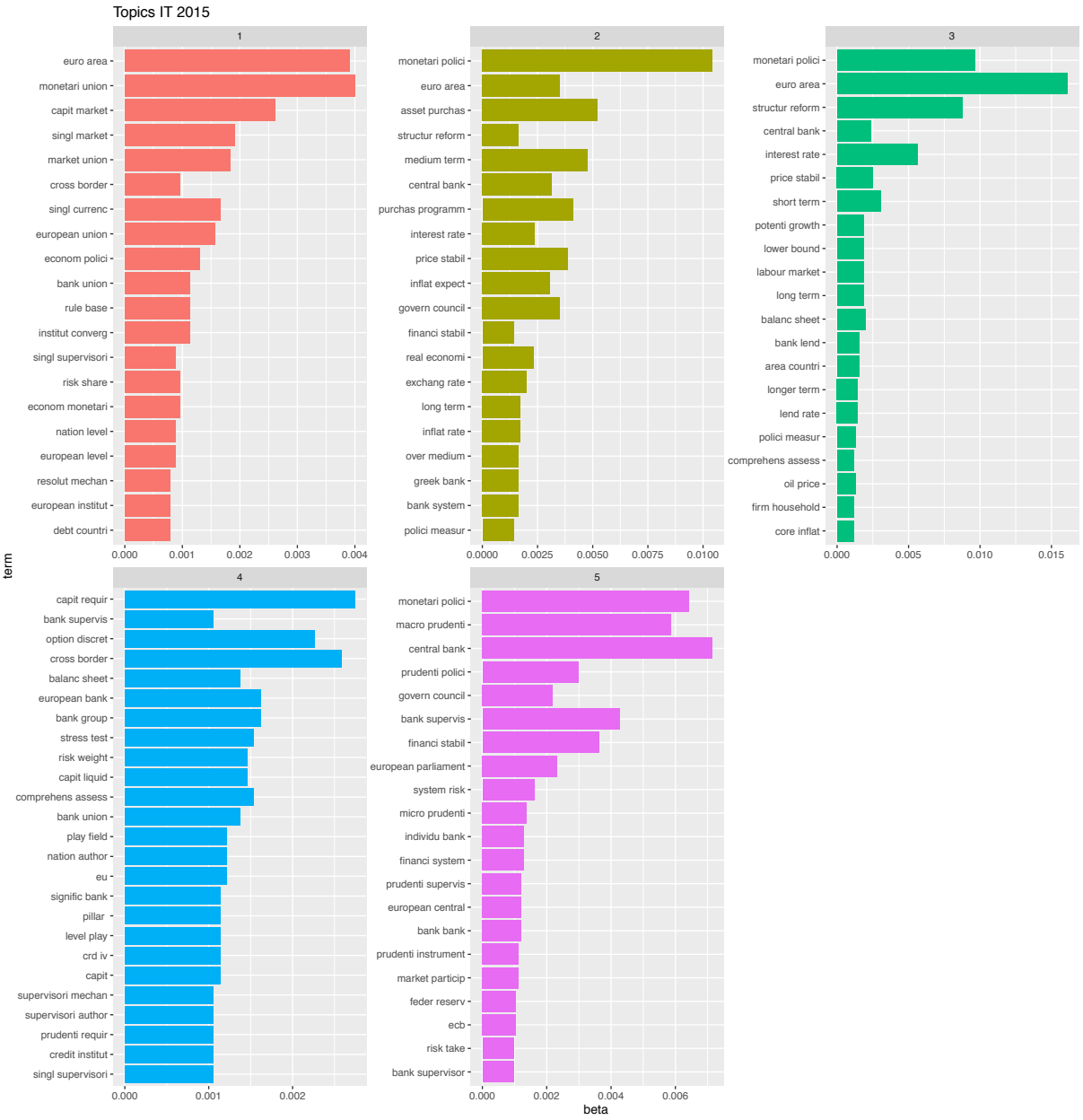

Topics PT 2014
1
euro area
bank union
bank sector
european bank
balanc sheet
financi crisi
cross border
financi stabil
comprehens assess
bank sovereign
monetari union
financi integr
european level
bank balanc
resolut bank
border bank
total
resolut fund
econom monetari
singl resolut
financi fragment
0.000
0.002
0.004
0.006
2
monetari polici
euro area
interest rate
central bank
low inflat
balanc sheet
yield curv
price stabil
inflat rate
inflat expect
asset purchas
non standard
govern council
forward guidanc
financi market
standard measur
exchang rate
rate
short term
purchas programm
polici stanc
feder reserv
0.0000
0.0025
0.0050
0.0075
3
monetari polici
euro area
bank union
emerg market
central bank
advanc economi
balanc sheet
financi crisi
long term
financi sector
structur reform
global financi
area countri
sovereign debt
unit state
econom growth
debt crisi
area bank
unconvent monetari
product growth
potenti growth
0.000
0.003
0.006
0.009
4
shadow bank
euro area
non bank
bank sector
financi stabil
financi system
financi institut
money market
bank system
financi market
real economi
short term
market fund
market base
invest fund
bank non
bank entiti
bank bank
total asset
financi regul
credit system
0.0000
0.0025
0.0050
0.0075
5
macro prudenti
prudenti polici
monetari polici
euro area
financi cycl
interest rate
central bank
system risk
financi stabil
price stabil
financi system
financi sector
capit requir
asset price
risk take
build up
risk weight
financi imbal
credit boom
leverag ratio
hous price
financi crise
boom bust
financi
0.000
0.005
0.010
0.015
term
beta

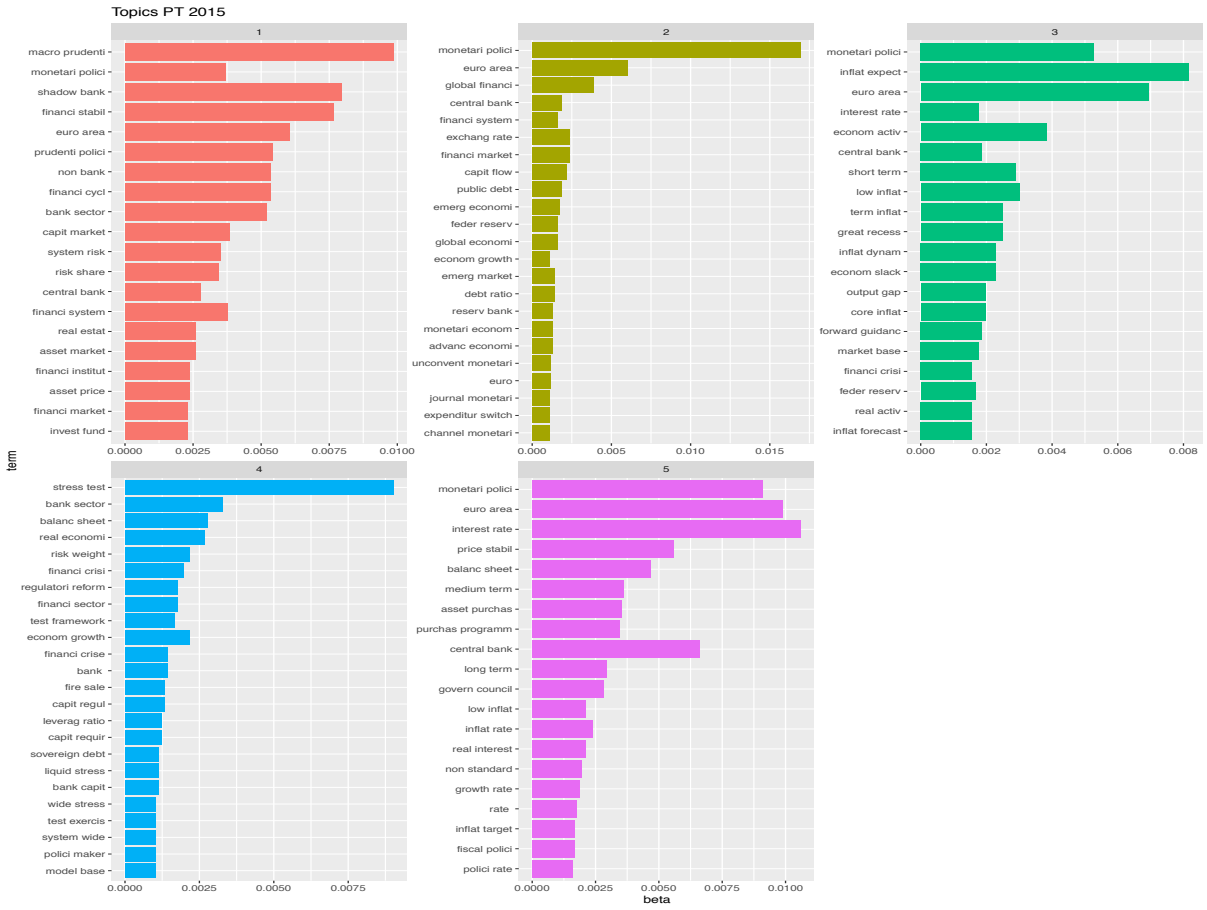

Topics BE 2015
1
euro area
monetari polici
inflat expect
asset purchas
medium term
govern council
lend rate
bank lend
purchas programm
credit eas
oil price
financi condit
firm household
close
below close
eas packag
low inflat
inflat rate
refinanc oper
expand asset
0.000 0.003 0.006 0.009
2
interest rate
euro area
monetari polici
potenti growth
long term
price stabil
balanc sheet
short term
term interest
output gap
real interest
bank sector
low level
low interest
bank union
growth rate
term rate
term growth
privat sector
financi crisi
credit condit
0.000 0.005 0.010 0.015
3
central bank
euro area
monetari polici
lower bound
price stabil
monetari union
balanc sheet
public sector
short term
monetari domin
zero lower
sector secur
econom slack
rate below
larg scale
fiscal polici
price level
inflat unemploy
fiscal author
unit state
purchas public
econom activ
0.000 0.005 0.010 0.015 0.020
4
central bank
govern council
govern bond
euro zone
low inflat
financi market
inflat rate
quantit eas
non perform
nation central
european central
bank abl
polici ecb
econom recoveri
ecb govern
swiss franc
rule european
monetari financ
low growth
irish govern
greek peopl
fulfil mandat
european level
crisi manag
council ecb
commod price
0.000 0.001 0.002 0.003
5
euro area
structur reform
monetari polici
labour market
long term
monetari union
area countri
long run
employe
fiscal polici
growth euro
creat condit
growth rate
run growth
market reform
growth expect
below
structur unemploy
small firm
product market
polici respons
exchang rate
0.000 0.005 0.010
beta
term

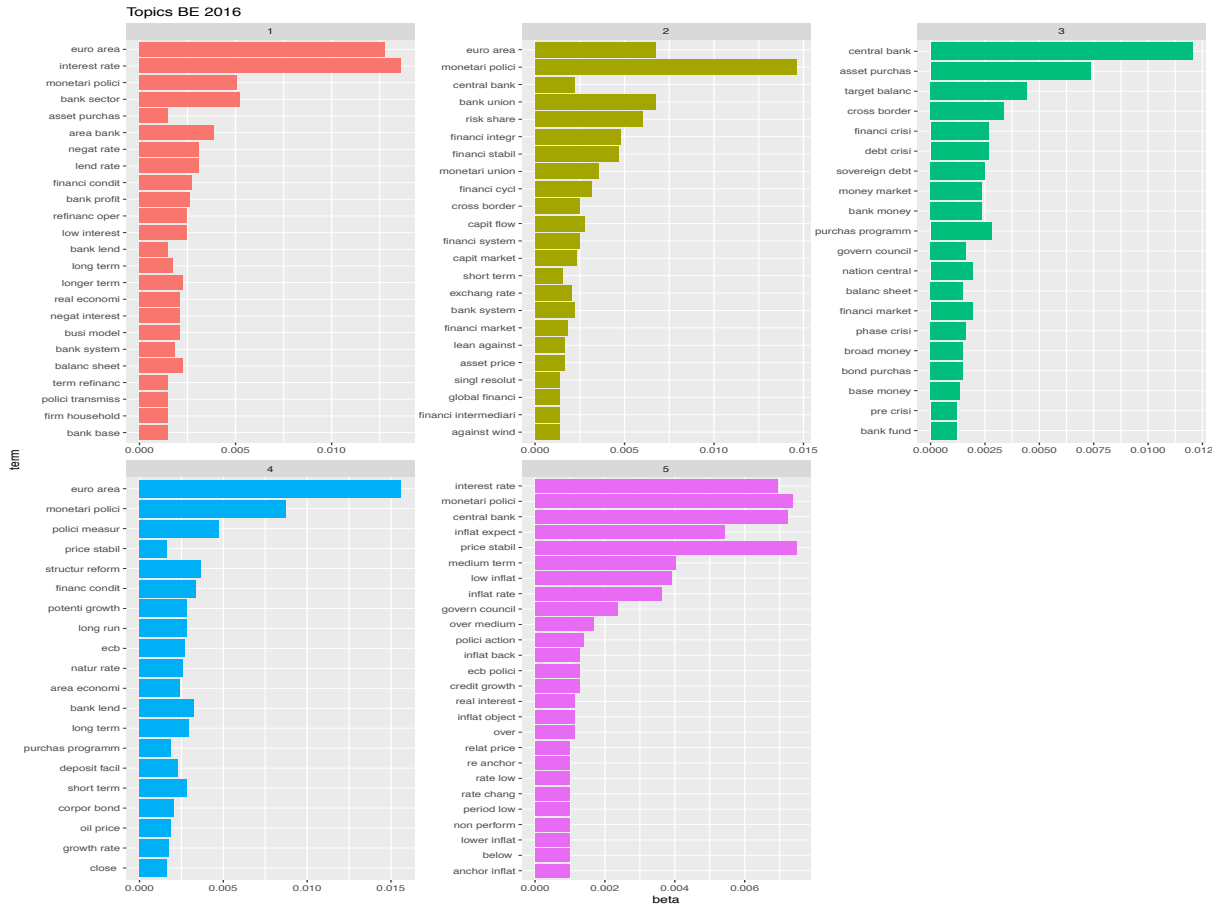

Topics DE 2015
1
financi market
supervis bank
european bank
long term
self regul
regul supervis
supervisori author
rule bank
european supervis
busi model
bank supervis
nation law
financi sector
bank industri
nation level
financi system
credit line
activ bank
german bank
european level
busi activ
0.000 0.001 0.002 0.003 0.004
2
euro area
central bank
risk profil
busi model
bank supervis
nation central
bank supervisor
singl supervisori
bank risk
bank group
granular data
european statist
play field
level play
govern council
financi stabil
escb statist
world statist
ssm supervis
follow up
financi statist
bank associ
balanc sheet
0.000 0.001 0.002 0.003 0.004
3
bank sector
trust bank
bank manag
long term
busi model
risk cultur
regain trust
lack trust
bank busi
risk appetit
polici maker
econom growth
regulatori supervisori
rebuild trust
inform incent
incent problem
asymmetr inform
area countri
affect bank
regul supervisor
bank function
bank activ
activ bank
0.000 0.002 0.004 0.006
4
srep decis
srep methodolog
capit requir
supervisori judgement
risk manag
bank system
bank supervis
quantit qualit
risk level
risk control
qualit supervis
nation discret
combin quantit
pillar
option nation
play field
level play
srep element
short term
safeti featur
risk bank
real economi
foster growth
ensur bank
0.000 0.001 0.002 0.003 0.004
5
interest rate
monetari polici
low interest
euro area
central bank
rate polici
financi market
purchas programm
polici measur
long term
structur reform
price stabil
low inflat
real interest
govern bond
econom growth
quantit eas
govern council
rate inflat
lower interest
key interest
inflat rate
0.000 0.005 0.010 0.015 0.020
beta
term

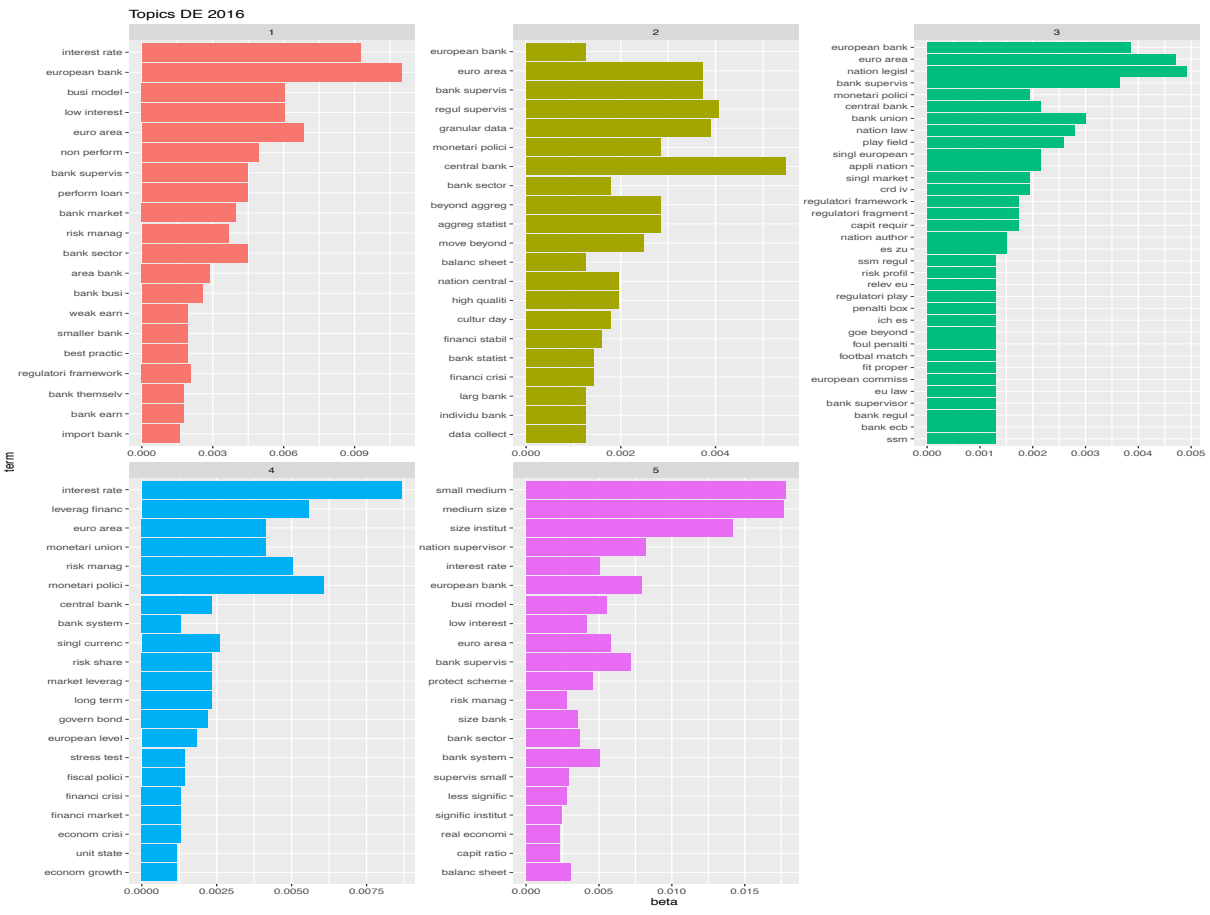

Topics FIN 2015
1
tier
price bubbl
possibl asset
oper op
insur compani
common equiti
bank insur
zone crisi
warn sector
up
take search
take note
strong risk
stabl countri
sovereign rate
solvenc margin
situat supervisori
search yield
russia jason
regard increas
refrain excess
rate countri
qualiti finland
problem mobil
posit finnish
phone sector
natur bank
low euro
level credit
largest bank
invest insur
household debt
hit euro
histor low
health warn
growth lend
generat net
finnish pension
face increas
except market
environ head
despit consecut
cut sovereign
custom finland
countri financi
countri central
concentr natur
compani strong
close eye
challeng interest
borrow cost
bank total
bank risk
bank includ
bank generat
bank bank
author incom
anna john
up
household
financi watchdog
0.000
0.005
0.010
0.015
0.020
2
financi sector
watchdog finnish
op subsidiari
economi contract
countri top
central bank
bank finnish
risk take
bank insur
yield prepar
watchdog finland
system concentr
supervisori account
risk addit
return output
report watchdog
report risk
rate environ
rate
pension fund
note finland
net interest
maker author
level central
junctur financi
interest incom
insur remain
impact custom
household indebted
hous market
growth averag
gdp yet
fund strong
finnish growth
finland vulner
financi system
eye economi
export hit
excess risk
estim annual
economi impact
economi financi
econom growth
deputi schedul
debt held
crise sever
contract quarter
contract chang
consecut yet
cite concern
capit ratio
capit posit
burst possibl
bubbl biannual
bear capac
averag cut
author junctur
agenc standard
adequaci rate
ad prolong
0.00
0.01
0.02
0.03
3
interest rate
market situat
financi supervisori
countri bank
capit adequaci
averag
risk take
zone interest
watchdog borrow
unforeseen risk
trade invest
total capit
stream vulner
state mean
slowdown russia
situat bring
sever countri
sector slowdown
requir particular
ratio averag
quarter yet
prepar burst
particip refrain
millennium half
market warn
market condit
margin countri
level export
lend overh
keep close
jason neeli
insur busi
indebted doubl
indebt
increas household
identifi vulner
household ad
hous debt
half hous
finland largest
finland deputi
face countri
credit agenc
consequ bank
condit ad
concentr eu
bank crise
bank anna
author bank
annual econom
addit identifi
ad risk
rate
0.00
0.01
0.02
0.03
0.04
4
supervisori author
return level
bank sector
asset price
watchdog
common
finnish bank
watchdog countri
vulner requir
vulner chang
vigil econom
take possibl
system face
statement overal
standard poor
st bank
separ estim
sector report
sector maintain
risk finland
risk erod
risk bear
report household
remain profit
remain health
remain below
prolong low
poor cite
polici maker
overal capit
output level
op st
mobil phone
mean consequ
insur solvenc
increas risk
increas challeng
includ low
held indebt
head financi
finnish gdp
finland financi
except econom
econom polici
econom financi
econom contract
domest economi
crisi problem
cost histor
contract consecut
compani face
chang market
busi financi
buffer keep
biannual report
bank remain
adequ capit
account statement
john
finland
capit
top lender
sector regard
euro zone
0.000
0.005
0.010
0.015
0.020
5
yet return
subsidiari bank
rate increas
low interest
lender oper
finland bank
financi market
equiti tier
bank finland
finnish bank
financi watchdog
top lender
sector regard
euro zone
warn financi
vulner financi
struggl domest
strong financi
sector particip
sector capit
schedul report
risk includ
reli trade
regard risk
ratio healthi
rate outlook
rate lead
profit despit
particular vigil
overh hous
outlook negat
negat stabl
market low
maintain adequ
low bank
lead growth
increas
incom stream
incom reli
includ op
healthi
growth remain
finnish economi
finland finnish
finland economi
eu state
erod struggl
earn qualiti
doubl millennium
debt concentr
consecut econom
concern finnish
chang except
capit buffer
capac insur
bubbl bank
bring unforeseen
below averag
bank system
bank separ
adequaci ratio
except
0.00
0.01
0.02
0.03
term
beta

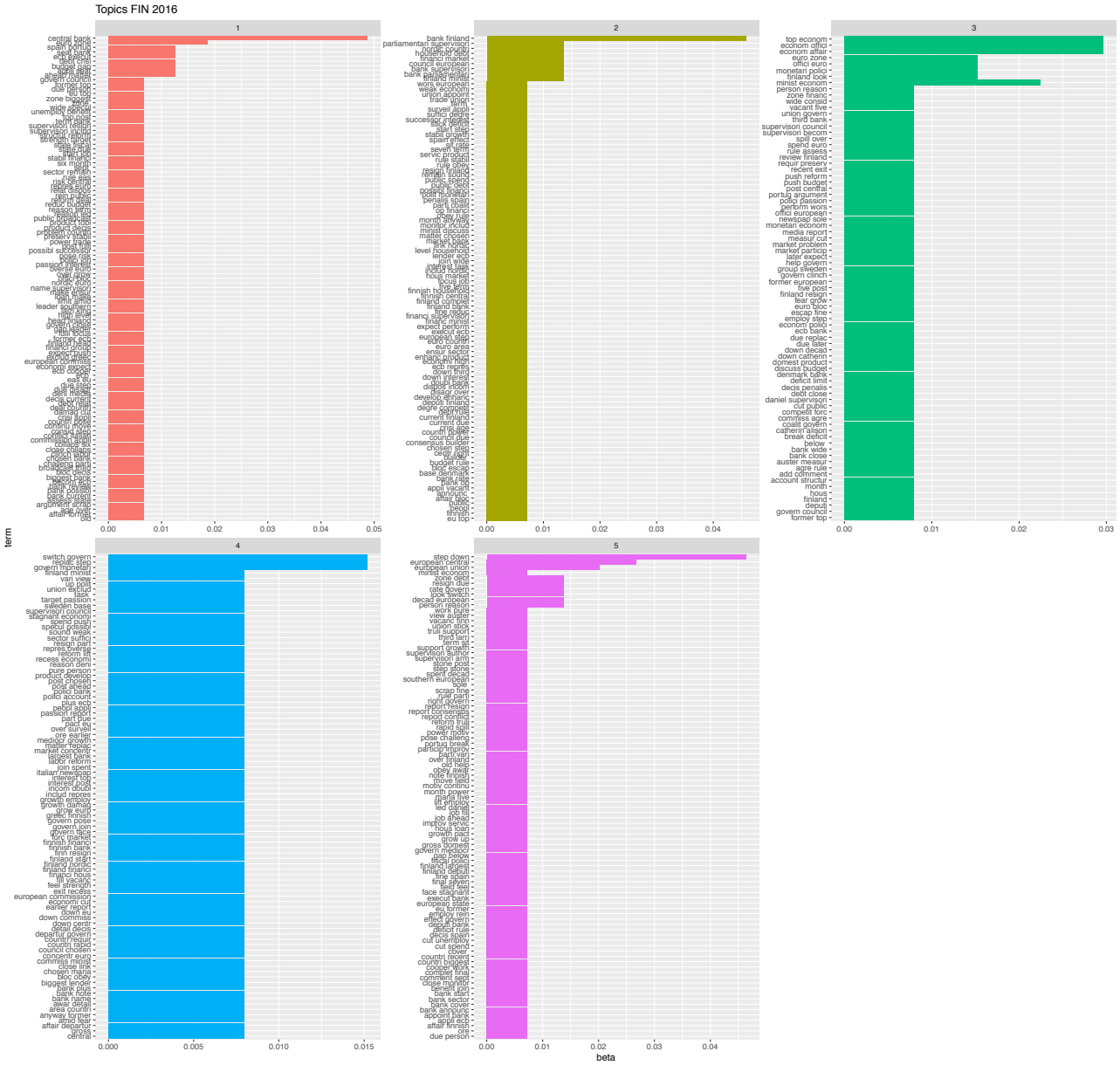

Topics FR 2015
1
euro area
monetari polici
monetari union
central bank
fiscal polici
interest rate
price stabil
bank union
govern council
singl currenc
long term
structur reform
european commiss
econom monetari
area countri
unit state
greek govern
european level
econom polici
oil price
0.000
0.005
0.010
2
european parliament
credit institut
central bank
main
ecb eu
european central
ecb
compet author
parliament european
european
bank main
econom monetari
parliament
intern audit
direct supervis
chair supervisori
ssm regul
regul eu
monetari affair
re
0.000
0.002
0.004
0.006
3
euro area
bank supervis
comprehens assess
european bank
capit requir
bank union
bank system
nation supervisor
busi model
bank sector
singl supervisori
balanc sheet
supervisori author
non perform
play field
financi crisi
signific institut
area bank
stress test
financi stabil
0.000
0.002
0.004
0.006
term
4
interest rate
govern bond
financi market
financi system
risk manag
low interest
greek bank
central clear
stress test
financi stabil
system risk
financi integr
risk weight
cross border
market particip
ccp risk
resolut fund
sovereign exposur
regulatori framework
data gap
0.0000
0.0005
0.0010
0.0015
0.0020
5
euro area
monetari polici
exchang rate
central bank
interest rate
lower bound
short term
intern currenc
global financi
asset purchas
financi market
financi crisi
purchas programm
financi stabil
financi integr
econom
capit market
capit flow
unit state
intern financi
global economi
0.000
0.002
0.004
0.006
beta

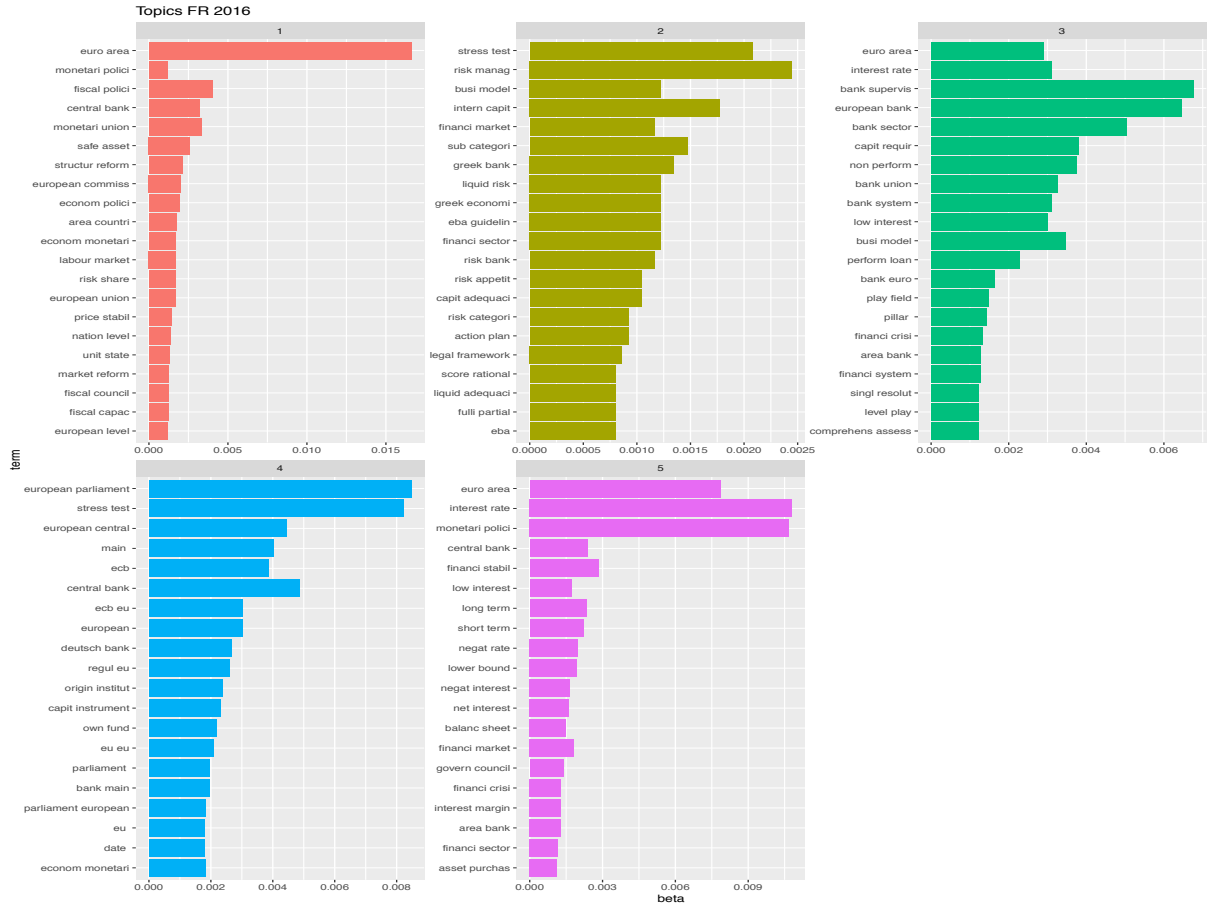

Topics IT 2015
1
euro area
monetari union
capit market
singl market
market union
cross border
singl currenc
european union
econom polici
bank union
rule base
institut converg
singl supervisori
risk share
econom monetari
nation level
european level
resolut mechan
european institut
debt countri
0.000
0.001
0.002
0.003
0.004
2
monetari polici
euro area
asset purchas
structur reform
medium term
central bank
purchas programm
interest rate
price stabil
inflat expect
govern council
financi stabil
real economi
exchang rate
long term
inflat rate
over medium
greek bank
bank system
polici measur
0.0000
0.0025
0.0050
0.0075
0.0100
3
monetari polici
euro area
structur reform
central bank
interest rate
price stabil
short term
potenti growth
lower bound
labour market
long term
balanc sheet
bank lend
area countri
longer term
lend rate
polici measur
comprehens assess
oil price
firm household
core inflat
0.000
0.005
0.010
0.015
4
capit requir
bank supervis
option discret
cross border
balanc sheet
european bank
bank group
stress test
risk weight
capit liquid
comprehens assess
bank union
play field
nation author
eu
signific bank
pillar
level play
crd iv
capit
supervisori mechan
supervisori author
prudenti requir
credit institut
singl supervisori
0.000
0.001
0.002
5
monetari polici
macro prudenti
central bank
prudenti polici
govern council
bank supervis
financi stabil
european parliament
system risk
micro prudenti
individu bank
financi system
prudenti supervis
european central
bank bank
prudenti instrument
market particip
feder reserv
ecb
risk take
bank supervisor
0.000
0.002
0.004
0.006
beta
term

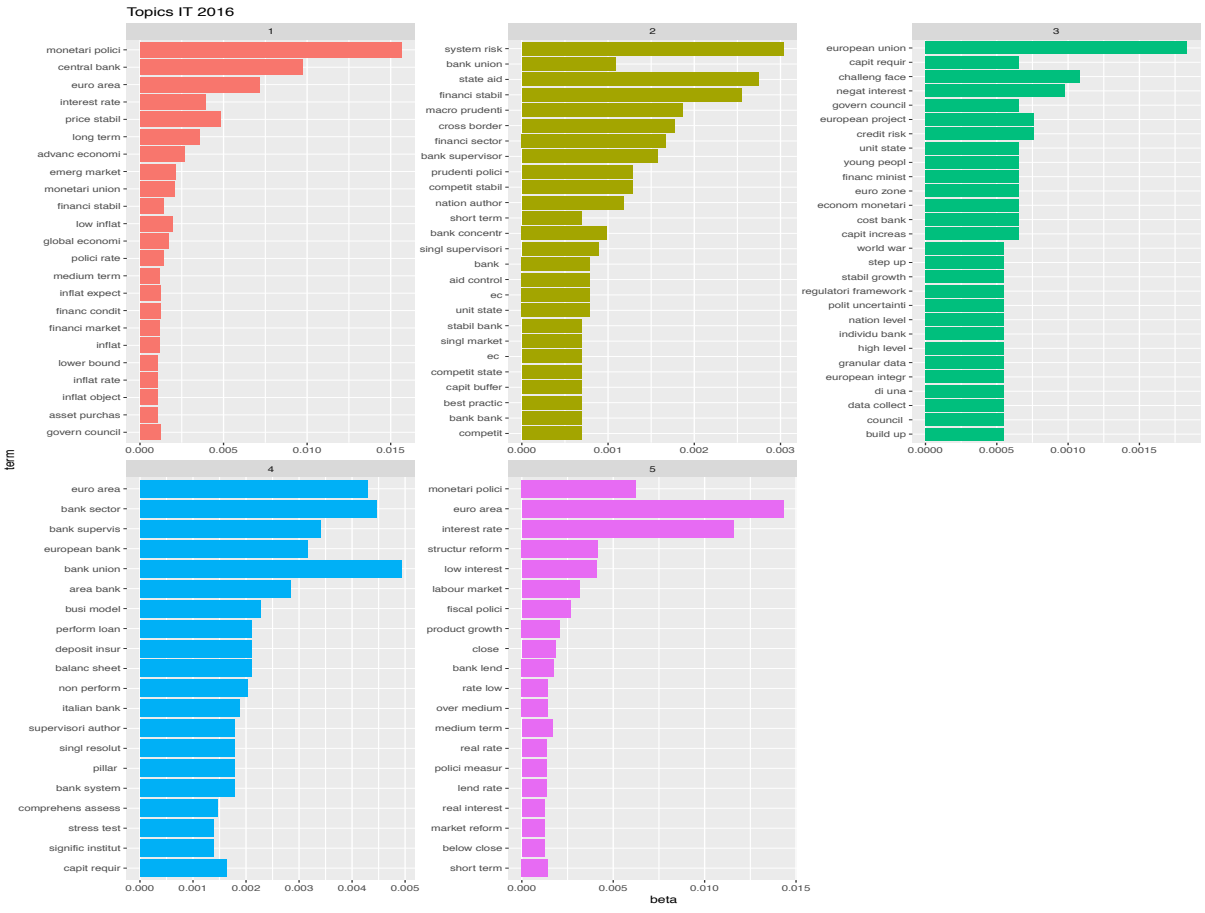

Topics PT 2015
1
macro prudenti
monetari polici
shadow bank
financi stabil
euro area
prudenti polici
non bank
financi cycl
bank sector
capit market
system risk
risk share
central bank
financi system
real estat
asset market
financi institut
asset price
financi market
invest fund
0.0000
0.0025
0.0050
0.0075
0.0100
2
monetari polici
euro area
global financi
central bank
financi system
exchang rate
financi market
capit flow
public debt
emerg economi
feder reserv
global economi
econom growth
emerg market
debt ratio
reserv bank
monetari econom
advanc economi
unconvent monetari
euro
journal monetari
expenditur switch
channel monetari
0.000
0.005
0.010
0.015
3
monetari polici
inflat expect
euro area
interest rate
econom activ
central bank
short term
low inflat
term inflat
great recess
inflat dynam
econom slack
output gap
core inflat
forward guidanc
market base
financi crisi
feder reserv
real activ
inflat forecast
0.000
0.002
0.004
0.006
0.008
4
stress test
bank sector
balanc sheet
real economi
risk weight
financi crisi
regulatori reform
financi sector
test framework
econom growth
financi crise
bank
fire sale
capit regul
leverag ratio
capit requir
sovereign debt
liquid stress
bank capit
wide stress
test exercis
system wide
polici maker
model base
0.0000
0.0025
0.0050
0.0075
5
monetari polici
euro area
interest rate
price stabil
balanc sheet
medium term
asset purchas
purchas programm
central bank
long term
govern council
low inflat
inflat rate
real interest
non standard
growth rate
rate
inflat target
fiscal polici
polici rate
0.0000
0.0025
0.0050
0.0075
0.0100
beta
term

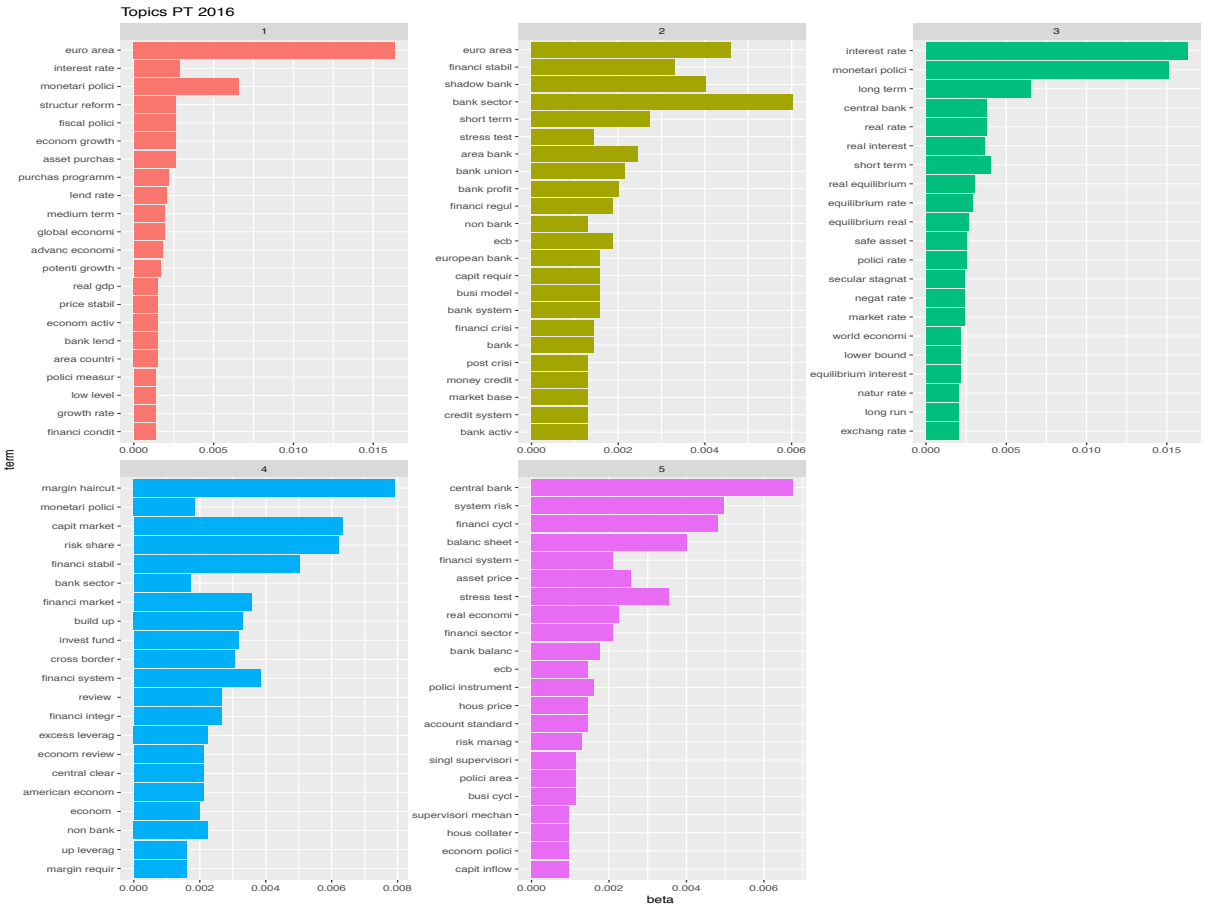

Topics BE 2016
1
euro area
interest rate
monetari polici
bank sector
asset purchas
area bank
negat rate
lend rate
financi condit
bank profit
refinanc oper
low interest
bank lend
long term
longer term
real economi
negat interest
busi model
bank system
balanc sheet
term refinanc
polici transmiss
firm household
bank base
0.000
0.005
0.010
2
euro area
monetari polici
central bank
bank union
risk share
financi integr
financi stabil
monetari union
financi cycl
cross border
capit flow
financi system
capit market
short term
exchang rate
bank system
financi market
lean against
asset price
singl resolut
global financi
financi intermediari
against wind
0.000
0.005
0.010
0.015
3
central bank
asset purchas
target balanc
cross border
financi crisi
debt crisi
sovereign debt
money market
bank money
purchas programm
govern council
nation central
balanc sheet
financi market
phase crisi
broad money
bond purchas
base money
pre crisi
bank fund
0.0000
0.0025
0.0050
0.0075
0.0100
0.012
4
euro area
monetari polici
polici measur
price stabil
structur reform
financ condit
potenti growth
long run
ecb
natur rate
area economi
bank lend
long term
purchas programm
deposit facil
short term
corpor bond
oil price
growth rate
close
0.000
0.005
0.010
0.015
5
interest rate
monetari polici
central bank
inflat expect
price stabil
medium term
low inflat
inflat rate
govern council
over medium
polici action
inflat back
ecb polici
credit growth
real interest
inflat object
over
relat price
re anchor
rate low
rate chang
period low
non perform
lower inflat
below
anchor inflat
0.000
0.002
0.004
0.006
beta
term

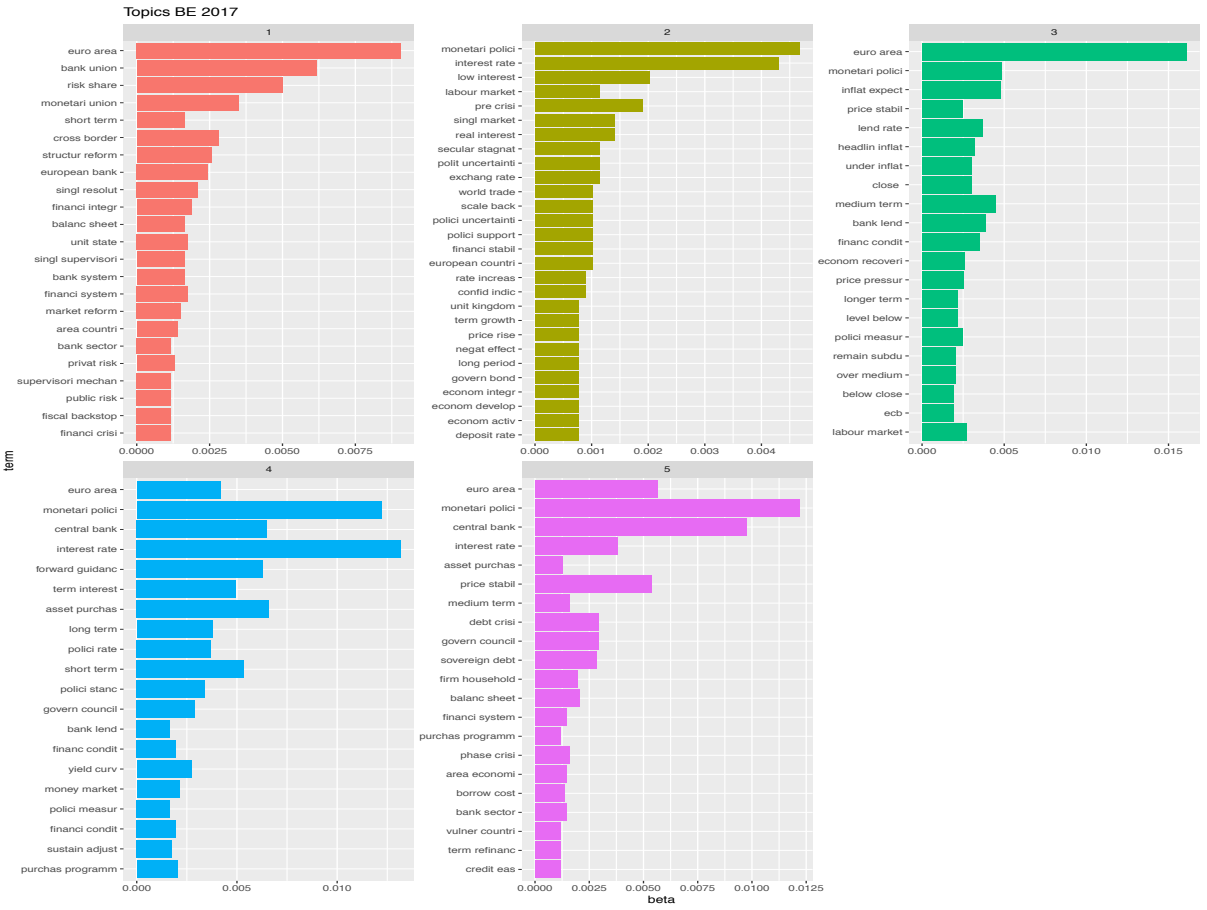

Topics DE 2016
1
interest rate
european bank
busi model
low interest
euro area
non perform
bank supervis
perform loan
bank market
risk manag
bank sector
area bank
bank busi
weak earn
smaller bank
best practic
regulatori framework
bank themselv
bank earn
import bank
0.000
0.003
0.006
0.009
2
european bank
euro area
bank supervis
regul supervis
granular data
monetari polici
central bank
bank sector
beyond aggreg
aggreg statist
move beyond
balanc sheet
nation central
high qualiti
cultur day
financi stabil
bank statist
financi crisi
larg bank
individu bank
data collect
0.000
0.002
0.004
3
european bank
euro area
nation legisl
bank supervis
monetari polici
central bank
bank union
nation law
play field
singl european
appli nation
singl market
crd iv
regulatori framework
regulatori fragment
capit requir
nation author
es zu
ssm regul
risk profil
relev eu
regulatori play
penalti box
ich es
goe beyond
foul penalti
footbal match
fit proper
european commiss
eu law
bank supervisor
bank regul
bank ecb
ssm
0.000
0.001
0.002
0.003
0.004
0.005
4
interest rate
leverag financ
euro area
monetari union
risk manag
monetari polici
central bank
bank system
singl currenc
risk share
market leverag
long term
govern bond
european level
stress test
fiscal polici
financi crisi
financi market
econom crisi
unit state
econom growth
0.0000
0.0025
0.0050
0.0075
5
small medium
medium size
size institut
nation supervisor
interest rate
european bank
busi model
low interest
euro area
bank supervis
protect scheme
risk manag
size bank
bank sector
bank system
supervis small
less signific
signific institut
real economi
capit ratio
balanc sheet
0.000
0.005
0.010
0.015
beta
term

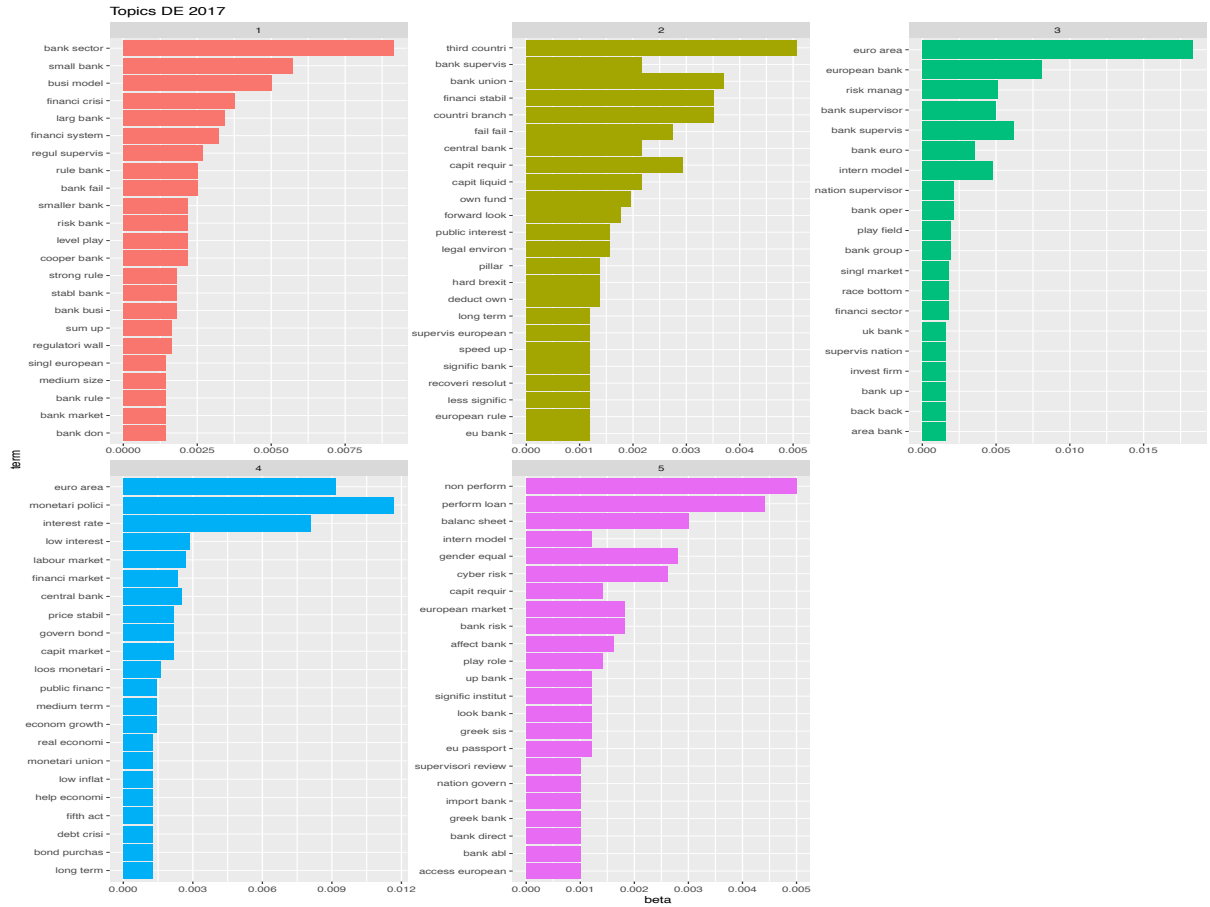

Topics FIN 2016
1
2
3
4
5
term
beta

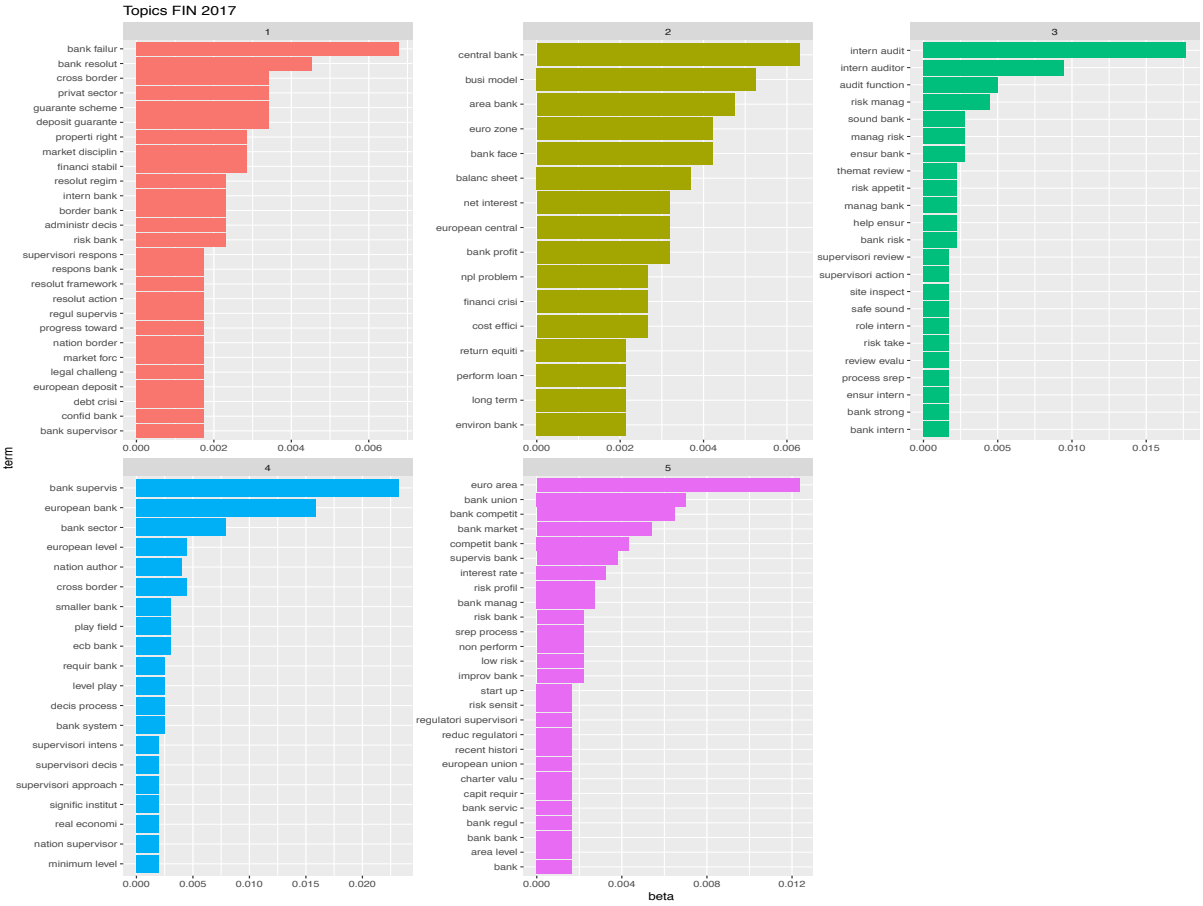

Topics FR 2016
1
euro area
monetari polici
fiscal polici
central bank
monetari union
safe asset
structur reform
european commiss
econom polici
area countri
econom monetari
labour market
risk share
european union
price stabil
nation level
unit state
market reform
fiscal council
fiscal capac
european level
0.000
0.005
0.010
0.015
2
stress test
risk manag
busi model
intern capit
financi market
sub categori
greek bank
liquid risk
greek economi
eba guidelin
financi sector
risk bank
risk appetit
capit adequaci
risk categori
action plan
legal framework
score rational
liquid adequaci
fulli partial
eba
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
3
euro area
interest rate
bank supervis
european bank
bank sector
capit requir
non perform
bank union
bank system
low interest
busi model
perform loan
bank euro
play field
pillar
financi crisi
area bank
financi system
singl resolut
level play
comprehens assess
0.000
0.002
0.004
0.006
4
european parliament
stress test
european central
main
ecb
central bank
ecb eu
european
deutsch bank
regul eu
origin institut
capit instrument
own fund
eu eu
parliament
bank main
parliament european
eu
date
econom monetari
0.000
0.002
0.004
0.006
0.008
5
euro area
interest rate
monetari polici
central bank
financi stabil
low interest
long term
short term
negat rate
lower bound
negat interest
net interest
balanc sheet
financi market
govern council
financi crisi
interest margin
area bank
financi sector
asset purchas
0.000
0.003
0.006
0.009
term
beta

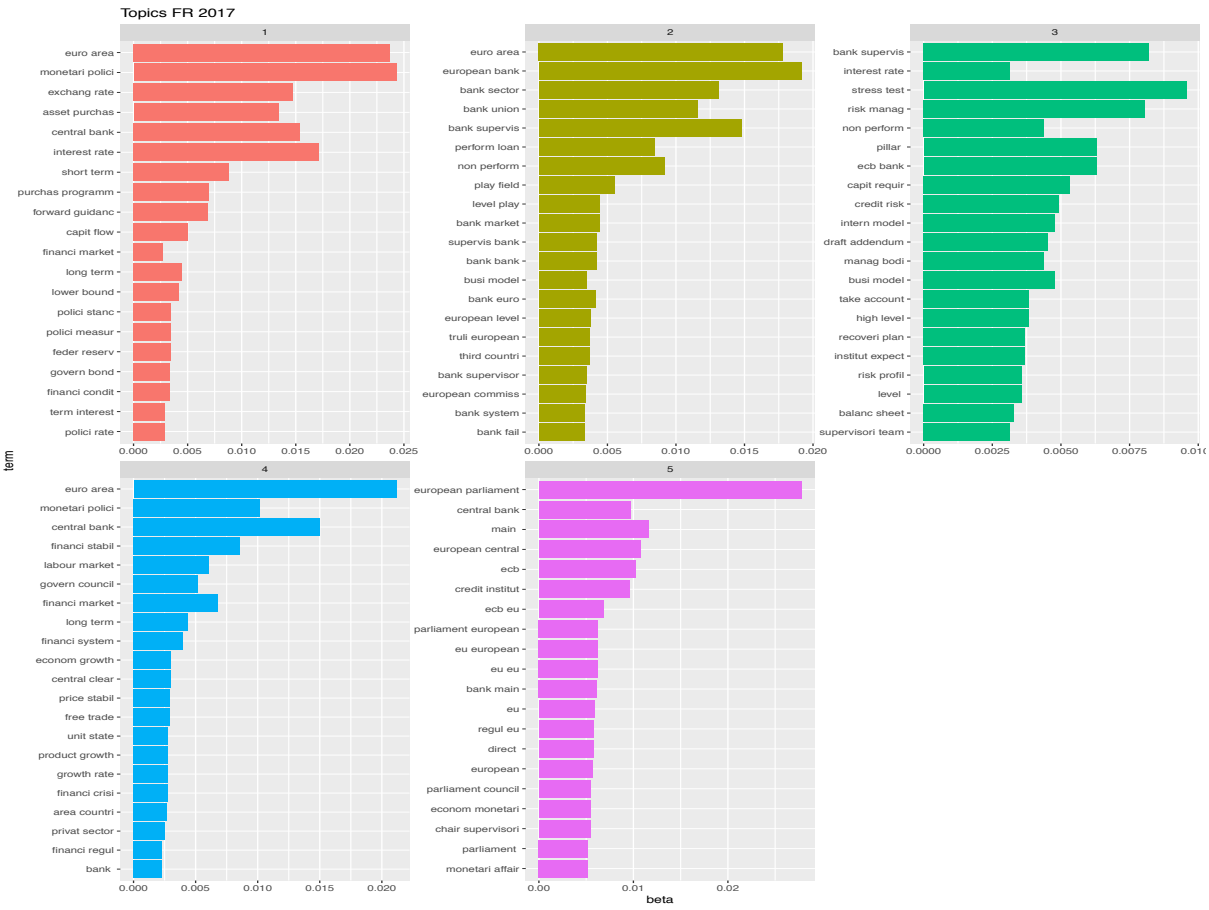

Topics IT 2016
1
monetari polici
central bank
euro area
interest rate
price stabil
long term
advanc economi
emerg market
monetari union
financi stabil
low inflat
global economi
polici rate
medium term
inflat expect
financ condit
financi market
inflat
lower bound
inflat rate
inflat object
asset purchas
govern council
0.000
0.005
0.010
0.015
2
system risk
bank union
state aid
financi stabil
macro prudenti
cross border
financi sector
bank supervisor
prudenti polici
competit stabil
nation author
short term
bank concentr
singl supervisori
bank
aid control
ec
unit state
stabil bank
singl market
ec
competit state
capit buffer
best practic
bank bank
competit
0.000
0.001
0.002
0.003
3
european union
capit requir
challeng face
negat interest
govern council
european project
credit risk
unit state
young peopl
financ minist
euro zone
econom monetari
cost bank
capit increas
world war
step up
stabil growth
regulatori framework
polit uncertainti
nation level
individu bank
high level
granular data
european integr
di una
data collect
council
build up
0.0000
0.0005
0.0010
0.0015
4
euro area
bank sector
bank supervis
european bank
bank union
area bank
busi model
perform loan
deposit insur
balanc sheet
non perform
italian bank
supervisori author
singl resolut
pillar
bank system
comprehens assess
stress test
signific institut
capit requir
0.000
0.001
0.002
0.003
0.004
0.005
5
monetari polici
euro area
interest rate
structur reform
low interest
labour market
fiscal polici
product growth
close
bank lend
rate low
over medium
medium term
real rate
polici measur
lend rate
real interest
market reform
below close
short term
0.000
0.005
0.010
0.015
term
beta

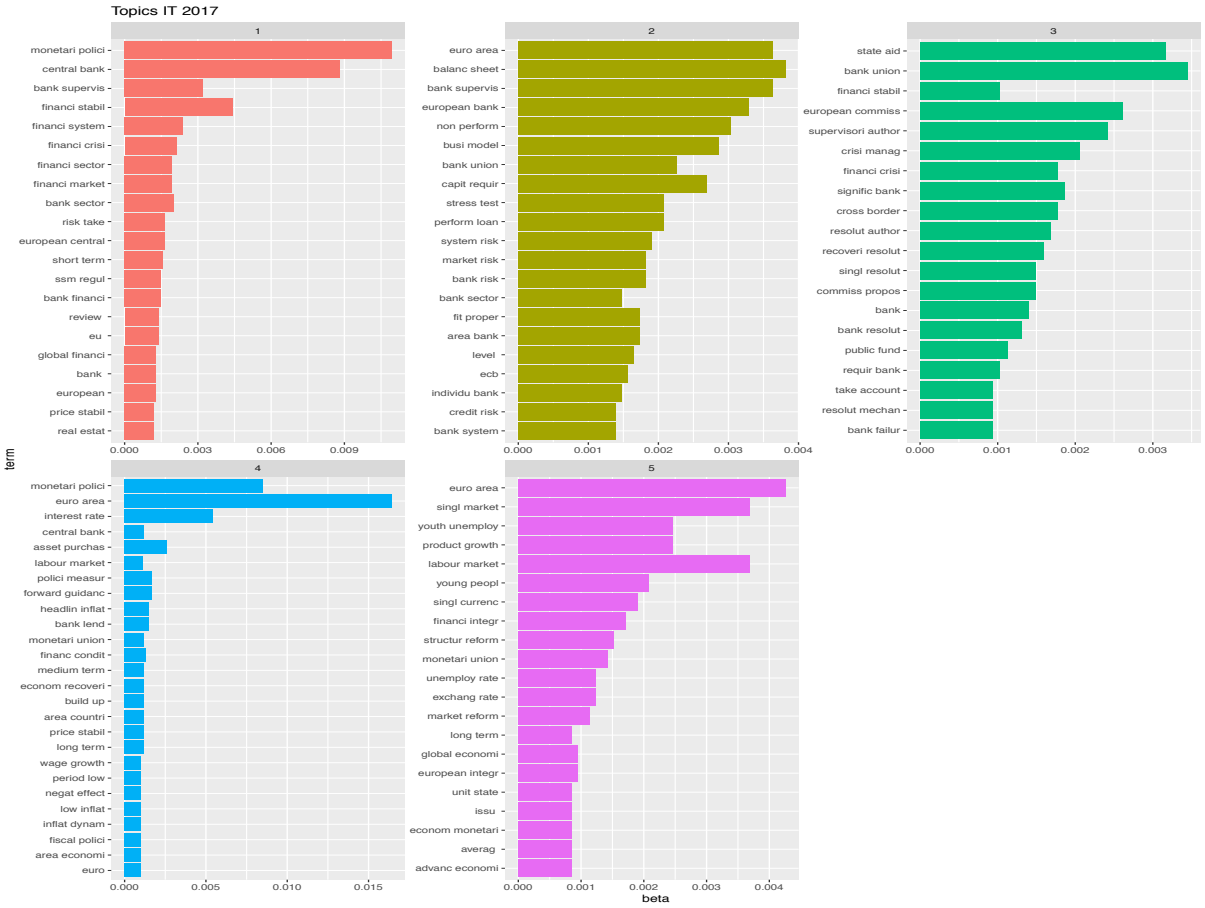

Topics PT 2016
1
euro area
interest rate
monetari polici
structur reform
fiscal polici
econom growth
asset purchas
purchas programm
lend rate
medium term
global economi
advanc economi
potenti growth
real gdp
price stabil
econom activ
bank lend
area countri
polici measur
low level
growth rate
financi condit
0.000
0.005
0.010
0.015
2
euro area
financi stabil
shadow bank
bank sector
short term
stress test
area bank
bank union
bank profit
financi regul
non bank
ecb
european bank
capit requir
busi model
bank system
financi crisi
bank
post crisi
money credit
market base
credit system
bank activ
0.000
0.002
0.004
0.006
3
interest rate
monetari polici
long term
central bank
real rate
real interest
short term
real equilibrium
equilibrium rate
equilibrium real
safe asset
polici rate
secular stagnat
negat rate
market rate
world economi
lower bound
equilibrium interest
natur rate
long run
exchang rate
0.000
0.005
0.010
0.015
term
4
margin haircut
monetari polici
capit market
risk share
financi stabil
bank sector
financi market
build up
invest fund
cross border
financi system
review
financi integr
excess leverag
econom review
central clear
american econom
econom
non bank
up leverag
margin requir
0.000
0.002
0.004
0.006
0.008
5
central bank
system risk
financi cycl
balanc sheet
financi system
asset price
stress test
real economi
financi sector
bank balanc
ecb
polici instrument
hous price
account standard
risk manag
singl supervisori
polici area
busi cycl
supervisori mechan
hous collater
econom polici
capit inflow
0.000
0.002
0.004
0.006
beta

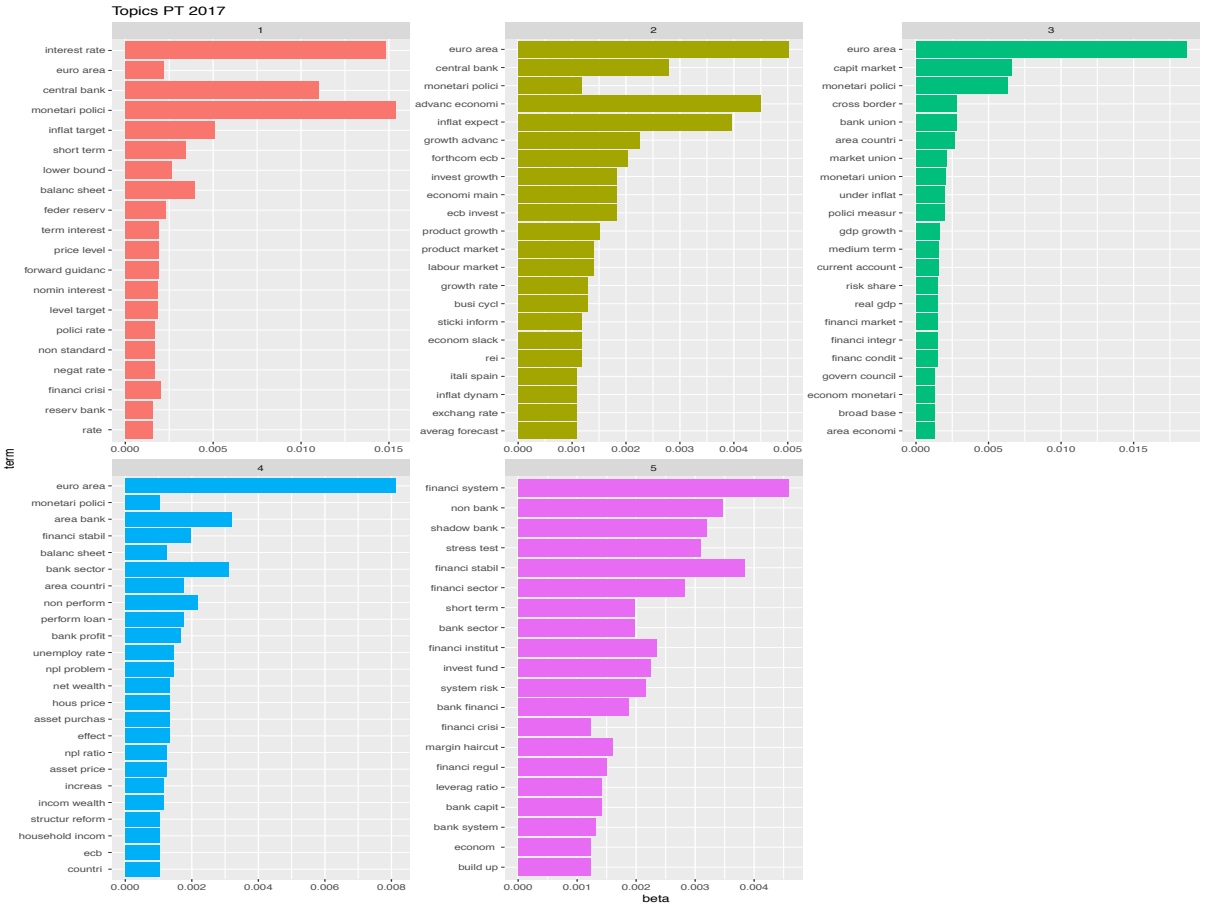

Topics BE 2017
1
euro area
bank union
risk share
monetari union
short term
cross border
structur reform
european bank
singl resolut
financi integr
balanc sheet
unit state
singl supervisori
bank system
financi system
market reform
area countri
bank sector
privat risk
supervisori mechan
public risk
fiscal backstop
financi crisi
0.0000
0.0025
0.0050
0.0075
2
monetari polici
interest rate
low interest
labour market
pre crisi
singl market
real interest
secular stagnat
polit uncertainti
exchang rate
world trade
scale back
polici uncertainti
polici support
financi stabil
european countri
rate increas
confid indic
unit kingdom
term growth
price rise
negat effect
long period
govern bond
econom integr
econom develop
econom activ
deposit rate
0.000
0.001
0.002
0.003
0.004
3
euro area
monetari polici
inflat expect
price stabil
lend rate
headlin inflat
under inflat
close
medium term
bank lend
financ condit
econom recoveri
price pressur
longer term
level below
polici measur
remain subdu
over medium
below close
ecb
labour market
0.000
0.005
0.010
0.015
4
euro area
monetari polici
central bank
interest rate
forward guidanc
term interest
asset purchas
long term
polici rate
short term
polici stanc
govern council
bank lend
financ condit
yield curv
money market
polici measur
financi condit
sustain adjust
purchas programm
0.000
0.005
0.010
5
euro area
monetari polici
central bank
interest rate
asset purchas
price stabil
medium term
debt crisi
govern council
sovereign debt
firm household
balanc sheet
financi system
purchas programm
phase crisi
area economi
borrow cost
bank sector
vulner countri
term refinanc
credit eas
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
term
beta


Topics DE 2017
1
bank sector
small bank
busi model
financi crisi
larg bank
financi system
regul supervis
rule bank
bank fail
smaller bank
risk bank
level play
cooper bank
strong rule
stabl bank
bank busi
sum up
regulatori wall
singl european
medium size
bank rule
bank market
bank don
0.0000
0.0025
0.0050
0.0075
2
third countri
bank supervis
bank union
financi stabil
countri branch
fail fail
central bank
capit requir
capit liquid
own fund
forward look
public interest
legal environ
pillar
hard brexit
deduct own
long term
supervis european
speed up
signific bank
recoveri resolut
less signific
european rule
eu bank
0.000
0.001
0.002
0.003
0.004
0.005
3
euro area
european bank
risk manag
bank supervisor
bank supervis
bank euro
intern model
nation supervisor
bank oper
play field
bank group
singl market
race bottom
financi sector
uk bank
supervis nation
invest firm
bank up
back back
area bank
0.000
0.005
0.010
0.015
4
euro area
monetari polici
interest rate
low interest
labour market
financi market
central bank
price stabil
govern bond
capit market
loos monetari
public financ
medium term
econom growth
real economi
monetari union
low inflat
help economi
fifth act
debt crisi
bond purchas
long term
0.000
0.003
0.006
0.009
0.012
5
non perform
perform loan
balanc sheet
intern model
gender equal
cyber risk
capit requir
european market
bank risk
affect bank
play role
up bank
signific institut
look bank
greek sis
eu passport
supervisori review
nation govern
import bank
greek bank
bank direct
bank abl
access european
0.000
0.001
0.002
0.003
0.004
0.005
term
beta


Topics FIN 2017
1
bank failur
bank resolut
cross border
privat sector
guarante scheme
deposit guarante
properti right
market disciplin
financi stabil
resolut regim
intern bank
border bank
administr decis
risk bank
supervisori respons
respons bank
resolut framework
resolut action
regul supervis
progress toward
nation border
market forc
legal challeng
european deposit
debt crisi
confid bank
bank supervisor
0.000
0.002
0.004
0.006
2
central bank
busi model
area bank
euro zone
bank face
balanc sheet
net interest
european central
bank profit
npl problem
financi crisi
cost effici
return equiti
perform loan
long term
environ bank
0.000
0.002
0.004
0.006
3
intern audit
intern auditor
audit function
risk manag
sound bank
manag risk
ensur bank
themat review
risk appetit
manag bank
help ensur
bank risk
supervisori review
supervisori action
site inspect
safe sound
role intern
risk take
review evalu
process srep
ensur intern
bank strong
bank intern
0.000
0.005
0.010
0.015
4
bank supervis
european bank
bank sector
european level
nation author
cross border
smaller bank
play field
ecb bank
requir bank
level play
decis process
bank system
supervisori intens
supervisori decis
supervisori approach
signific institut
real economi
nation supervisor
minimum level
0.000
0.005
0.010
0.015
0.020
5
euro area
bank union
bank competit
bank market
competit bank
supervis bank
interest rate
risk profil
bank manag
risk bank
srep process
non perform
low risk
improv bank
start up
risk sensit
regulatori supervisori
reduc regulatori
recent histori
european union
charter valu
capit requir
bank servic
bank regul
bank bank
area level
bank
0.000
0.004
0.008
0.012
term
beta


Topics FR 2017
1
euro area
monetari polici
exchang rate
asset purchas
central bank
interest rate
short term
purchas programm
forward guidanc
capit flow
financi market
long term
lower bound
polici stanc
polici measur
feder reserv
govern bond
financi condit
term interest
polici rate
0.000
0.005
0.010
0.015
0.020
0.025
2
euro area
european bank
bank sector
bank union
bank supervis
perform loan
non perform
play field
level play
bank market
supervis bank
bank bank
busi model
bank euro
european level
truli european
third countri
bank supervisor
european commiss
bank system
bank fail
0.000
0.005
0.010
0.015
0.020
3
bank supervis
interest rate
stress test
risk manag
non perform
pillar
ecb bank
capit requir
credit risk
intern model
draft addendum
manag bodi
busi model
take account
high level
recoveri plan
institut expect
risk profil
level
balanc sheet
supervisori team
0.0000
0.0025
0.0050
0.0075
4
euro area
monetari polici
central bank
financi stabil
labour market
govern council
financi market
long term
financi system
econom growth
central clear
price stabil
free trade
unit state
product growth
growth rate
financi crisi
area countri
privat sector
financi regul
bank
0.000
0.005
0.010
0.015
0.020
5
european parliament
central bank
main
european central
ecb
credit institut
ecb eu
parliament european
eu european
eu eu
bank main
eu
regul eu
direct
european
parliament council
econom monetari
chair supervisori
parliament
monetari affair
0.00
0.01
0.02
term
beta


Topics IT 2017
1
monetari polici
central bank
bank supervis
financi stabil
financi system
financi crisi
financi sector
financi market
bank sector
risk take
european central
short term
ssm regul
bank financi
review
eu
global financi
bank
european
price stabil
real estat
0.000
0.003
0.006
0.009
2
euro area
balanc sheet
bank supervis
european bank
non perform
busi model
bank union
capit requir
stress test
perform loan
system risk
market risk
bank risk
bank sector
fit proper
area bank
level
ecb
individu bank
credit risk
bank system
0.000
0.001
0.002
0.003
0.004
3
state aid
bank union
financi stabil
european commiss
supervisori author
crisi manag
financi crisi
signific bank
cross border
resolut author
recoveri resolut
singl resolut
commiss propos
bank
bank resolut
public fund
requir bank
take account
resolut mechan
bank failur
0.000
0.001
0.002
0.003
4
monetari polici
euro area
interest rate
central bank
asset purchas
labour market
polici measur
forward guidanc
headlin inflat
bank lend
monetari union
financ condit
medium term
econom recoveri
build up
area countri
price stabil
long term
wage growth
period low
negat effect
low inflat
inflat dynam
fiscal polici
area economi
euro
0.000
0.005
0.010
0.015
5
euro area
singl market
youth unemploy
product growth
labour market
young peopl
singl currenc
financi integr
structur reform
monetari union
unemploy rate
exchang rate
market reform
long term
global economi
european integr
unit state
issu
econom monetari
averag
advanc economi
0.000
0.001
0.002
0.003
0.004
beta
term


Topics PT 2017
1
interest rate
euro area
central bank
monetari polici
inflat target
short term
lower bound
balanc sheet
feder reserv
term interest
price level
forward guidanc
nomin interest
level target
polici rate
non standard
negat rate
financi crisi
reserv bank
rate
0.000
0.005
0.010
0.015
2
euro area
central bank
monetari polici
advanc economi
inflat expect
growth advanc
forthcom ecb
invest growth
economi main
ecb invest
product growth
product market
labour market
growth rate
busi cycl
sticki inform
econom slack
rei
itali spain
inflat dynam
exchang rate
averag forecast
0.000
0.001
0.002
0.003
0.004
0.005
3
euro area
capit market
monetari polici
cross border
bank union
area countri
market union
monetari union
under inflat
polici measur
gdp growth
medium term
current account
risk share
real gdp
financi market
financi integr
financ condit
govern council
econom monetari
broad base
area economi
0.000
0.005
0.010
0.015
4
euro area
monetari polici
area bank
financi stabil
balanc sheet
bank sector
area countri
non perform
perform loan
bank profit
unemploy rate
npl problem
net wealth
hous price
asset purchas
effect
npl ratio
asset price
increas
incom wealth
structur reform
household incom
ecb
countri
0.000
0.002
0.004
0.006
0.008
5
financi system
non bank
shadow bank
stress test
financi stabil
financi sector
short term
bank sector
financi institut
invest fund
system risk
bank financi
financi crisi
margin haircut
financi regul
leverag ratio
bank capit
bank system
econom
build up
0.000
0.001
0.002
0.003
0.004
beta
term

References

Acosta, M. (2015). FOMC Responses to Calls for Transparency.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3-32.

Bailey, A., & Schonhardt-Bailey, C. (2008). Does deliberation matter in FOMC monetary policymaking? The Volcker Revolution of 1979. *Political Analysis*, 16(4), 404-427.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636.

Bank of England, "One Bank Research Agenda". Discussion paper, February 2015.

Bank of International Settlements (2010). Basel III: A global regulatory framework for more resilient banks and banking systems. Basel, CH.
Bellman, R. E. (2015). Adaptive control processes: a guided tour. *Princeton university press*.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics*, *120*(1), 387-422.

Bholat, D. (2015). Big data and central banks. *Big Data & Society*, *2*(1), 2053951715579469.

Bholat, D., Brookes, J., Cai, C., Grundy, K., & Lund, J. (2017). Sending firm messages: text mining letters from PRA supervisors to banks and building societies they regulate. Working Paper.

Bholat, D. M., Hansen, S., Santos, P. M., & Schonhardt-Bailey, C. (2015). Text mining for central banks. Working Paper.

Black, F. (1995). Interest rates as options. *The Journal of Finance*, 50(5), 1371-1376.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Blinder A.S., & Ehrmann M., & Fratzscher M., & De Haan J., & Jansen D., 2008. "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence," *Journal of Economic Literature*, American Economic Association, vol. 46(4), pages 910-45, December.

Born, B., Ehrmann, M., & Fratzscher, M. (2014). Central bank communication on financial stability. *The Economic Journal*, 124(577), 701-734.

Bruno, G. (2016). Text mining and sentiment extraction in central bank documents. In *2016 IEEE International Conference on Big Data* (pp. 1700-1708). IEEE.

Bulíř, A., Čihák, M., & Jansen, D. J. (2013). What drives clarity of central bank communication about inflation?. *Open Economies Review*, 24(1), 125-145.

Campbell, J. R., Evans, C. L., Fisher, J. D., & Justiniano, A. (2012). Macroeconomic effects of Federal Reserve forward guidance. *Brookings Papers on Economic Activity*, *2012*(1), 1-80.

Carretta, A., Farina, V., Fiordelisi, F., Schwizer, P., & Lopes, F. S. S. (2015). Don't Stand So Close to Me: The role of supervisory style in banking stability. *Journal of Banking & Finance*, 52, 180-188.

Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. BoE Staff working paper N°674.

Cœuré B. (2017, November). 'Policy analysis with big data'. Conference on Economic and Financial Regulation in the Era of Big Data. Banque de France, Paris.

Conti, A. M., Neri, S., & Nobili, A. (2015). Why is inflation so low in the euro area?.

Ehrmann, M., & Fratzscher, M. (2007). Communication by central bank committee members: different strategies, same effectiveness?. *Journal of Money, Credit and Banking*, 39(2-3), 509-541.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*,*346*(6210), 1243089.

European Union, (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union* (OJ), 59(1-88), 294.

European Central Bank (2014), The ECB's Forward Guidance. In Monthly Bulletin, April 2014. Frankfurt, DE.

European Central Bank (2015), ECB annual report on supervisory activities, March 2015. Frankfurt, DE.

European Central Bank (2016), ECB Banking Supervision: SSM supervisory priorities 2017 , December 2016. Frankfurt, DE.

European Central Bank (2017a), Guidance to banks on non-per- forming loans, March 2017. Frankfurt, DE.

European Central Bank (2017b), Addendum to the ECB Guidance to Banks on Non- performing Loans: Prudential Provisioning Backstop for Non-performing Exposures. Frankfurt, DE.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management,* 23(3), 157-214.

Gaffeo, E., & Mazzocchi, R. (2019). “The price is right”: using auction theory to enhance competition in the NPL market. *Journal of Banking Regulation*, 1-9.

Galardo, M., & Guerrieri, C. (2017). The Effects of Central Bank's Verbal Guidance: Evidence from the ECB. Bank of Italy Temi di Discussione (Working Paper) No, 1129.

Goldsmith-Pinkham, P., Hirtle, B., & Lucca, D. O. (2016). Parsing the content of bank supervision.

Gürkaynak, R.S., Sack, B., Swanson, E., 2005. Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *Int. J. Cen. Bank.* 1 (1), May.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, S114-S133.

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2), 801-870.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer, New York, NY.

Hayo, B., & Neuenkirch, M. (2015). Self-monitoring or reliance on media reporting: How do financial market participants process central bank news?. *Journal of Banking & Finance*, *59*, 27-37.

Hernández-Murillo, R., & Shell, H. (2014). The rising complexity of the FOMC statement. *Economic Synopses*, (23).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Joseph, A. (2019). Staff Working Paper No. 784 Shapley regressions: a framework for statistical inference on machine

learning models.

Kahveci, E., & Odabaş, A. (2016). Central Banks' Communication Strategy and Content Analysis of Monetary Policy Statements: The Case of Fed, ECB and CBRT. *Procedia-Social and Behavioral Sciences*, *235*, 618-629.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128-147.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

Meade, E. E., & Acosta, M. (2015). Hanging on every word: Semantic analysis of the FOMC's postmeeting statement (No. 2015-09-30). *Board of Governors of the Federal Reserve System* (US).

Moessner, R., Jansen, D. J., & de Haan, J. (2017). Communication about future policy rates in theory and practice: A survey. *Journal of Economic Surveys*, 31(3), 678-711.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, 42(3), 1314-1324.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

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

Nyman, R., Gregory, D., Kapadia, S., Ormerod, P., Tuckett, D., & Smith, R. (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment. BoE, mimeo.

Oosterloo, S., & de Haan, J. (2004). Central banks and financial stability: a survey. *Journal of Financial Stability*, 1(2), 257-273.

Praet, P. (2013). Forward Guidance and the ECB. Forward Guidance: Perspectives from Central Bankers, Scholars and Market Participants, Vox eBook.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14-46.

Rosa, C., & Verga, G. (2007). On the consistency and effectiveness of central bank communication: Evidence from the ECB. *European Journal of Political Economy*, 23(1), 146-175.

Schmeling, M., & Wagner, C. (2019). Does central bank tone move asset prices?.

Schonhardt-Bailey, C. (2013). *Deliberating American monetary policy: a textual analysis.* MIT Press.

Shirota, Y., Yano, Y., Hashimoto, T., & Sakura, T. (2015). Monetary policy topic extraction by using LDA: Japanese monetary policy of the second ABE cabinet term. In 2015 *IIAI 4th International Congress on Advanced Applied Informatics* (pp. 8-13). IEEE.

Smaghi, L. B. (2009). Conventional and unconventional monetary policy. Speech at the Center for Monetary and Banking Studies, Geneva, 28.

Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *The Journal of Economic Perspectives*, *15*(4), 101-115.

Stough, R., & McBride, D. (2014). Big data and US public policy. Review of Policy Research, Vol. 31, No. 4, pp. 339-342.

Swanson, E. T. (2017). Measuring the effects of Federal Reserve forward guidance and asset purchases on financial markets (No. w23311). National Bureau of Economic Research.

Takeda, Y., & Keida, M. (2018). Central bank communication strategies: A computer-based narrative analysis of the Bank of Japan's Governor Kuroda. *Hawks and Doves: Deeds and Words*, 137.

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

Tobback, E., Nardelli, S., & Martens, D. (2017). Between hawks and doves: measuring central bank communication.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3- 27.

Woodford, M. (2005). *Central bank communication and policy effectiveness* (No. w11898). National Bureau of Economic Research.

Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero-lower bound. *Journal of Money, Credit and Banking*, 48(2-3), 253-291.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics* (Vol. 16, No. 13, p. S8). BioMed Central.

Classifying Firms with Text Mining

ABSTRACT

Statistics on the births, deaths and survival rates of firms are crucial pieces of information, as they enter as an input in the computation of GDP, the identification of each sector's contribution to the economy, and the assessment of gross job creation and destruction rates. Official statistics on firm demography are made available only several months after data collection and storage, however. Furthermore, unprocessed and untimely administrative data can lead to a misrepresentation of the life-cycle stage of a firm. In this paper we implement an automated version of Eurostat's algorithm aimed at distinguishing true startup endeavors from the resurrection of pre-existing but apparently defunct firms. The potential gains from combining machine learning, natural language processing and econometric tools for pre-processing and analyzing granular data are exposed, and a machine learning method predicting reactivations of deceptively dead firms is proposed.

Keywords: Business Demography; Classification; Text Mining.

JEL classification: C01, C52, C53, C55, C80, G33, L11, L25, L26, M13, R11.

1 Introduction

The expression “big data” is usually associated with the recent explosion in the availability and richness of potentially relevant data due to the technological progress in collecting and storing them (Diebold, 2003). From this point of view, the archives and data repositories of public administrations are incredibly rich data sources, although their effective exploitation in economic and statistical research has traditionally encountered several hurdles (Stough and McBride, 2014).

The first one is strictly related to the qualitative nature of the information at hand. Because of their massive scale and peculiar features, qualitative and textual data cannot be handled by humans in a fast and costless manner. Moreover, the noise and lack of standardization characterizing them is an issue for both primary and preprocessed data (Schintler and Kulkarni, 2014), highlighting the need for advanced pre-processing techniques. Finally, these databases are often too large to be stored in standard electronic storehouses and analyzed with standard statistical and econometric techniques. As an example, in text mining a corpus of documents is usually represented as a *document term matrix*, in which each column is associated to a word (the regressor) and the cells indicate the presence of a word in a document. Such a matrix can easily reach a width of thousands of different features, meaning that it becomes exceptionally hard to handle it without sophisticated statistical tools.

The second problem in using public archives as data sources has to do with the timeliness of their dissemination in a usable form. Although economists and statisticians are usually asked to produce short, medium and long run forecasts on economic variables like GDP, inflation, unemployment and other descriptive variables regarding business dynamics and sectorial evolution, their estimates are often based on incomplete and uncertain sources (Castle *et al*., 2009). Official statistics derived from administrative data are published only several months – if not years – after their collection and processing. This calls for the need to combine known data and forecasts in an appropriate way. As emphasized by Einav and Levin (2014), exciting improvements can be made by exploiting information coming from private and administrative records in combination with machine learning (ML), natural language processing (NLP) and econometric tools.

The focus of this work is on business demography and the starting point is the Eurostat manual on business demography statistics (2007).

Eurostat releases statistics[1] on businesses’ births and deaths at the NUTS3 level (that is, at the province level) with a time lag of two years due to technical reasons resulting from the very definition of births and deaths. Indeed, a business is not really dead if it is reactivated within two years after the

[1] Population of active enterprises released by Eurostat fully covers activities from section B to N of NACE Rev.2; data for sections P, Q, R and S are provided on a voluntary basis; agriculture is excluded from the count, as well as banking. When computing births, holding companies are considered as ancillary activities while the set of active firms includes all sections of NACE Rev.2, including A and O.

cessation. It follows that, in order to have a definitive count of the number of deaths and births, one should wait two years with the consequence that reliable statistics on business demography are unavailable in real time. As a matter of fact, entrepreneurship plays a crucial role in GDP growth (Van Stel et al., 2005) and business statistics are ancillary to the computation of GDP and to the identification of each sector's contribution to the economy (Ahmad, 2008). The ambition of this work is to provide a method aimed at obtaining statistical information on business demography two years before the official ones. For this purpose, real time data and forecasts are combined.

As startups can be linked to pre-existing firms by some elements of continuity, for statistical reasons it is useful to distinguish startups from regenerated companies in order to prevent the creation of firms and the death of businesses from being misestimated. Moreover, a reliable count of demographic events is fundamental when public policies are introduced and their effects over entrepreneurship are evaluated. Nonetheless, classifying firms manually is expensive, time-consuming and the subjective judgement of human beings can lead to biased results. Consequently, the potential for exploiting statistical softwares and machine learning algorithms clearly emerges.

We implement an automated version of Eurostat's classification algorithm capable of detecting continuity linkages between firms in order to distinguish new births from creations which don't represent new combinations of production factors. The algorithm is applied to firms recorded at the business register of the Chamber of Commerce of Bolzano (Italy) and results are compared with the statistics released by Eurostat. Furthermore, one of the objectives of this paper is to present a method to predict whether a dormant or apparently dead firm will reactivate within two years. We exploit text mining and machine learning in order to process unstructured and noisy administrative data, to improve classification performances and to investigate demographic events characterizing firms during their lifecycle. Since bureaucracy and protocols for collecting and recording information evolve over time, as does technology, administrative data contain collection and registration errors, or they can be correct but poorly coherent. Non-conventional approaches are proposed with the aim to standardize data and to replace missing information. Text mining methods are applied to classify and code data in order to make them consistent within the dataset and capable of international comparison. Moreover, machine learning tools are investigated further for forecasting and assessing variables' relevance.

To the best of our knowledge, this work is the first attempt[2] to classify the economic sector and to

[2] The Italian National Institute of Statistics (ISTAT) provides a toolkit called RELAIS including a set of techniques for dealing with record linkage. It is possible to obtain a NACE code as an output inserting a keyword describing the economic activity performed by the firm as an input. A possible drawback of this tool is that it works just for the Italian version of the NACE code (namely, ATECO) and it can be automated hardly. Moreover, its classification performances are questionable. As an instance, we used the Italian word "ristorante" (meaning "restaurant") as an input and we obtained as an output the NACE code 49.1, associated to the activity of "passenger rail transport". This could be due to the fact that such a software works properly when the inserted keyword is matched precisely with a textual description contained in the NACE handbook, but it doesn't work with synonyms. As it will be

replace missing information on administrative data based on text mining. Instead, the earlier work by Roelands et al. (2017) is aimed at classifying firms' economic sectors by scraping information from their websites. In their application the best performing algorithm was found to be the naïve Bayes while in our case random forests perform better, as will be shown in the reminder of the paper. Recent works that are more related to our study are Gweon et al. (2017) and Ikudo et al. (2018). Both of them deal with automated occupational coding; similarly to us, Ikudo et al. (2018) conclude that the best classification algorithm is random forest, but in a frame with a relatively small number of classes.

This work is addressed to practitioners dealing with the elaboration of statistics on business demography and to academics investigating business dynamics. Indeed, the study of the antecedents of firms' survival and failure has been a central topic within management literature. Researchers have investigated thoroughly the reasons why firms fail or survive with the aim of assessing survival conditions. However, according to Josefy *et al*. (2017), some relevant grey areas should be still investigated since prior empirical papers studying the life and death of firms failed to take into account the apparently failed entities that are revitalized in new ones.

The rest of the paper is organized as follows. Section 2 provides a background on business demography. Section 3 describes the dataset and defines variables. Section 4 briefly introduces text mining and machine learning techniques adopted and discusses their classification performances with reference to our data. In Section 5 Eurostat's classification algorithm is implemented while in Section 6 the results previously obtained are combined with other data to build a training set; a random forest is estimated, and reactivations are forecasted. Section 7 concludes.

Due to Eurostat's effort aimed at promoting the production and use of harmonized data[3] on businesses demography, we believe our method can be applied to handle data collected in any European Union and OECD country.

2 Business Demography: Births and Deaths

Business entries and exits are recorded in business registers. Nevertheless, just a subset of firms' creations and deletions should be classified as *births* or *deaths* (Eurostat, 2007). Start-ups can be linked to pre-existing firms -still active or apparently dead- by some elements of continuity. For example, a new business might be run by an entrepreneur leading an apparently dead and formally different pre-existing firm conducting the same economic activity. According to Eurostat (2007) a

argued in the reminder of the paper, this is the reason why use textual descriptions provided by the entrepreneur to the business register to train the model.

3 See Eurostat (2010).

birth is "*the creation of a combination of production factors with the restriction that no other enterprises are involved in the event*". As a matter of fact, business registrations and deletions can be the consequences of a wide range of demographic events such as mergers, splits off, break-ups, changes of ownership (including successions), take-overs, joint ventures, and restructuring of enterprises or groups. Consequently, when a business death is recorded, it can be the signal of an event occurring within the life-cycle of the firm rather than at the end of it. For statistical reasons it is important to identify such events in order to prevent the creation or death of businesses from being misestimated as they are ancillary to the production of other statistics and to the evaluation of implemented public policies.

Reference units and classification criteria

This work focuses on two main events: enterprise birth and death. The reference statistical unit is the enterprise, as defined by System of National Accounts and International Standard of Industrial Classifications (1993) and the suggested data sources are administrative records such as businesses registers, tax registers, statistical surveys and social security data. In practice, business registers are the primary and preferred sources in light of the harmonization process involving EU members (see Eurostat, 2010).

The identification of births and deaths is strictly connected with the definition of active enterprises, that is, those that had either turnover or employment at any time[4] during the reference period (Eurostat, 2007). In practice, implementing the algorithm, the reference population can be defined according to the information set concretely available and to researcher's preferences; nevertheless, if we consider the evolving role of physical assets and labour force (see Acemoglu and Restrepo, 2017) and the progressive informative power deterioration of accounting documents (Lev and Gu, 2016), the definition of an active firm proposed by Eurostat may appear anachronistic. In that spirit, since this document is mainly aimed at dealing with the issue of detecting elements of continuity among the reference statistical units, the theoretical definition of an active firm is beyond the scope of this paper. For our purposes, active firms are those which gave communication to the businesses register of the start of their economic activity, eventually specifying the economic activity code.

Analyzing demographic events, two aspects matter: i) the number of active firms before and after the event and ii) the degree of continuity between the new business and the pre-existing ones. As far as the first aspect is concerned, if an enterprise is active at time *t* and is not at time *t+1* (or

4 The specification "at any time during the reference period" is interpreted in such a way as to remove the need for the reference enterprises to still be active on 31.12.

viceversa) it doesn't necessarily follow that a business is dead (or a new firm is created). When a merger occurs, *n* firms that were active at time *t* coalesce in the only firm remaining at time *t+1*; in this case, we should not report *n* deaths and *1* birth. Logically, if an entrepreneur gives up his or her enterprise to an heir who modifies the company name and the legal form, the consequent business deletion from specific registers should not count as a death, just as the new record is not a birth.[5]

In order to establish in which cases births and deaths effectively occur, the degree of continuity among firms matters. The continuity is addressed considering three main features characterizing the firm: legal control, economic activity and location. Roughly speaking, a new recorded activation does not constitute a birth if it can be matched with a pre-existing business based on at least two out of three common features. Moreover, if a company changes its legal form from unlimited liability to limited, formally causing a deletion, the new recorded business presents continuity with a (formally death) pre-existing firm - even though it moves to a new location - to the extent in which the principal economic activity remains the same: so, it should not be classified as a birth. To conclude, the birth is defined by Eurostat (2007) as "*the creation of a new combination of factors of production and […] the creation of a new enterprise reference on the business registers*" while the death is the "*dissolution of a combination of factors of production*"; both events are supposed to concern just one enterprise, leading respectively to a record and a deletion from the registers. It follows that, whenever a break-up occurs and an enterprise splits its production factors into two or more new enterprises, the latter are not new births and the original one is not dead, even though it is deleted from the registers.

Working the proposed algorithm at the enterprise level, there is a plethora of demographic events involving local units that should be handled carefully. As well as the enterprise, the local unit can born, die and be transferred (Eurostat, 2010). Events occurring at the local level do not necessarily matter at the enterprise one and viceversa. Localizations carrying out economic activities on behalf of the legal entity that exhibits control over them can gain or lose their legal identity, without implying a birth or a death. Similarly, a local unit which is split from the initial entity and transferred to a new controlling subject does not represent a business birth -although it is recorded according to a new identification number or acquires its own legal identity- to the extent in which the aforementioned continuity rules hold.

It can also be the case that a business effectively ceases its activity without involving any other firms. In this case, although a deletion occurs as a consequence of the interruption of the economic activity, the business is definitely dead unless it is reactivated within two years. From the way in which the enterprise death is defined it follows that, in order to have an exact count of deaths and births, a theoretical time-lag of two years is faced. Bearing such a discrepancy in mind, one aim of

5 An accurate description of any possible event (with its proper classification) can be found in Eurostat (2007).

this work is to use ML tools to predict the extent in which a dormant or apparently dead firm is likely to reactivate within two years, in order to obtain a real-time description of demographic events characterizing a given area for the reference period.

Identifying births

According to the European Commission (1998), enterprise births are "*A count of the number of births of enterprises registered to the population concerned in the business register corrected for errors. [...] Births do not include entries into the population due to: mergers, break-ups, split-off or restructuring of a set of enterprises"*. Events involving more than one enterprise, as well as ancillary[6] activities, are excluded.

Eurostat (2007) provides a theoretical method, defined in five steps, aimed at identifying new births among the enterprise creations. The steps are the following: *i*) identification of enterprises that were active at time *t* over which the analysis is conducted; population of active enterprises at time *t-1* and *t-2* should be generated too; *ii*) identifying new activations between 01.01 and 31.12: that is, those enterprises that are present in the population of active enterprises at time *t* but were not active at *t-1*; *iii*) comparing new activations at time *t* with firms active at *t-2* in order to detect reactivations of dormant firms; *iv*) elimination of creations due to events other than births; new activations identified at steps *i* and *ii* are matched according to a pairwise process controlling for activity[7]-location, location-name and activity-name.[8] Reactivations defined at step *iii* are excluded; *v*) correction of errors. After the identification is performed by computer algorithms, further manual investigations are needed. To this extent, the aim of this document is to expose an efficient way of automatizing the implementation of the Eurostat's algorithm with R (with special attention paid to the continuity issue) minimizing the role of manual checks and, to finalize, refining the process by text mining techniques especially aimed at replacing missing information and correcting errors. The entire procedure should include local units as well.

Identifying deaths

Deaths are treated by Eurostat (2007) and the European Commission (1998) as complementary to births, that is "*A death amounts to the dissolutions of a combination of production*

6 Ancillary activities are defined based on ISIC Rev. 4 and NACE Rev. 2 codes (see Eurostat, 2007, chapter 5).

7 The economic activity is proxied by the 4-digit level of ISIC-NACE as in Eurostat (2007).

8 On the definition of that variable a certain flexibility is left to the researcher. In existing literature, expressions like controlling legal unity, legal form and name are used rather equivalently (Eurostat, 2007 and Ahmad, 2008). As will be argued throughout, appropriate proxies should be defined.

factors with the restrictions that no other enterprises are involved in the event. Deaths do not include exit into the population due to: mergers, take-overs, break-ups and restructuring of a set of enterprises." In practice, cessations are identified by comparing the population of enterprises active at time *t* (obtained at the first step of the algorithm run for births) with the population of active enterprises at time *t+1* and *t+2* in order to detect cessations and reactivations occurring within two calendar years, respectively. As for births, a pairwise matching process should be carried out in order to detect continuity links between ceased enterprises and other enterprises or local units according to the three already exposed criteria. The matching process is aimed at distinguishing apparent deaths from real ones. Assuming that the count of deaths (as well as that of births) is performed at the beginning of *t+1* with reference to *t,* it is clear that, although it is theoretically possible to provide real time estimates for new births, the count of deaths requires data about the population of active enterprises at time *t+1* and *t+2* not available at the end of *t*. For that reason, Eurostat provides data on births and deaths with a time lag of two years. The issue of producing provisional data on deaths will be addressed within the document.

3 Dataset Description and Matching Process

Our data spans from 2012 to 2017. Data come from two sources: one primary source, that is the business register of the Chamber of Commerce of Bolzano (Italy) containing a wide range of demographic information, and a secondary source represented by the proprietary database Aida-Bureau van Dijk which provided accounting data. In detail, we have historical demographic data about more than 200,000 firms that were recorded at the business register of Bolzano and with at least one legal unit in the province. Among those, more than 58,000 were active between 2012 and 2017. Table 1 and 2 describe the composition of the dataset.

Table 1: Legal form

	2012	2013	2014	2015	2016	2017
Cooperatives	949	981	1018	1058	1069	1064
Individual proprietorships	37723	37716	37030	37131	37001	36972
Institutions	9	9	9	9	9	183
Joint-stock companies	5756	5892	6164	6778	7214	8001
One-man joint-stock companies	2223	2480	2432	2413	2309	2260
Other forms	240	251	245	247	236	890
Private partnerships	11246	11093	11022	10927	10817	10566
Total	**58146**	**58422**	**57920**	**58563**	**58655**	**59936**

Notes: the table describes the composition of the dataset according to the firms' legal form. The rise in the number of observations in 2017 is partially due to the inclusion (for administrative reasons) in the business register of firms previously recorded elsewhere.

Table 2: Economic Sectors

	2012	2013	2014	2015	2016	2017
Agriculture, farming and fishing	17477	17466	17090	17023	16982	17139
Arts, sport and entertainment	489	486	486	487	496	541
Automotive, transport and storage	10416	10427	10362	10338	10289	10418
Banking and insurances	677	701	702	715	721	742
Education	177	189	195	202	203	264
Handworks	15	6	9	5	9	6
Information technology	925	964	980	1062	1101	1150
Manufacturing	4188	4121	4076	4092	4065	4039
Others activities and services	1804	1810	1780	1801	1806	1861
Procurement, supplies and extractions	679	830	993	1246	1296	1360
Public administration, defence, health and social protection	125	129	135	145	155	176
Real estate and constructions	9228	9240	9204	9291	9278	9396
Sciences and techniques	1948	2052	2092	2257	2358	2425
Tourism	8523	8606	8595	8730	8809	8982
NA	1475	1395	1221	1169	1087	1437
Total	**58146**	**58422**	**57920**	**58563**	**58655**	**59936**

Notes: the table describes the composition of the dataset according to the firms' sector. For each year just active firms are considered and only the activity of the main local unit is accounted for.

We know the detailed localization of each firm and local unit, its legal form and the tax number. We have information about any demographic event including date of registration, date of activation, date and cause of closure and deletion from the register as well as unstructured textual data describing further events involving the firm. The NACE[9] code is known for more than 150,000 firms, accurately describing which economic activity is (or was) performed and in which local unit; it sums up to more than 300,000 local units including branches and offices. Moreover, additional textual descriptions of the activities are directly provided by the firm to the register and further textual information for craft businesses are available[10]. As far as individuals or legal entities involved in the firm are concerned, the dataset contains information about the shareholders, including their share of ownerships, their demographic details and their current place of residence[11]. Finally, demographic information about corporate officers and members of boards and committees (including bankruptcy trustees), their role in the firm and their appointment period are available too.

The practical implementation of the algorithm is not as trivial as it may appear. The underlying idea of the matching process is a pairwise comparison involving for each firm any single local unit. Since the matching algorithm is based on detecting overlaps of at least two of the three variables mentioned in Section 2 between two (formally) different firms, a proxy for those variables needs to be defined, given the amount of available information. The starting approach will seek for exact matches between features belonging to pairs of statistical units. In order to exploit the potential of fast packages freely available on R[12], text mining techniques will be applied preprocessing data, correcting errors and replacing missing information[13].

In this work, the *location* of the firm, or more generally, of the local unit, is defined rather restrictively as the full address; in a more extensive interpretation, two firms can be thought to operate in the same location if they merely reside in the same town. From such an extensive definition a potential underestimation of the birth rate follows since more matches (resulting in more continuity linkages) can potentially be found.

Another possible criticism refers to the definition of *legal unit*. Basically, the legal unit coincides with the corporate name; unfortunately, the informative power of the name is very poor. In order to perform pairwise comparisons between all the combinations of firms, the need for similarity

9 The NACE code is called “Ateco” in Italy.

10 The strategy adopted to replace missing data or to enrich the existing ones using textual information will be explained subsequently.

11 In some cases, based on the way the surname is formed, we can also generate the last name of the consort. We implicitly assume the form “X-in-Y” to mean that the shareholder “X” is married to “Y”.

12 https://www.r-project.org/

13 Further details will be given in the following Sections.

measures would arise[14] and the algorithm would increase in computational complexity; so, detecting exact matches between the individuals or the legal entities controlling the firm based on univocal tax codes appears to be a viable strategy, once data are preprocessed and made consistent. Nevertheless, the definition of control should be clarified. Dealing with control two aspects matter: management and ownerships. As a matter of fact, the extent to which board members, managers and entrepreneurs effectively lead the firm depends on the size of the firm itself and on the role of shareholders and other non-proprietorship key-figures in the hierarchy. In the present implementation of the algorithm, the continuity between two firms is evaluated by verifying the overlapping presence of controlling individuals or entities. Identifying a controlling shareholder is a matter of setting a threshold for the share of ownership; the percentage of ownership beyond which a participation is considered to be relevant clearly implies practical consequences: the lower the threshold, the higher the number of individuals and entities involved in the matching process and the greater the likelihood that a continuity linkage will occur. To make it quite clear, a threshold of 5% for the share of ownership implies the great majority of shareholders to be compared across firms causing a relatively high rate of start-ups to be classified as "fake". As far as administrative figures are concerned, a set of controlling positions should be arbitrarily defined by the researcher. Table A1 in the appendix provides the full list of controlling positions adopted for the present research.

The matching process is aimed at verifying the continuity between a new firm and a preexisting one or between an apparently ceased firm and an active one. In concrete terms, statistical units are compared in pairs under the assumption that there is continuity if combinations of, at least, localization and economic activity, localization and control or control and economic activity coincide for two firms. Practically speaking, each row of the dataset represents the combination of a controlling individual or entity (preferably proxied by the tax code), a given local unit of a firm (that is, the address) and the economic activity performed there (a 4-digit code); by merging local units and controlling entities by row we are left with each row representing a localization, its activity and a univocal identifying the controlling entity; this is repeated for each possible combination of local unit, activity and natural or legal persons.

To make the matching process possible some preprocessing is needed. The next Section is devoted to introducing learning classification techniques aimed at replacing missing information and correcting errors. In order to classify the economic activity performed by the firm based on the textual description when the NACE code is missing, Linear discriminant analysis (LDA), Naïve Bayes (NB) and Random forests (RF) are compared. Different strategies for address correction and

14 The use of similarity indexes based on the company sign or the business name should be investigated. Text mining and document-term distances can be exploited as it is very unlikely that a firm remains active and retains the same company name after a demographic event.

standardization are proposed and useful R functions are briefly examined.

4 Text Mining for Classification, Error Correction and Missing Data Imputation

Tokenization and preprocessing

In this Section some machine learning techniques are briefly described and their use for data manipulation is suggested. Text mining coupled with learning algorithms can be useful for error correction and missing information replacement. Indeed, to the extent in which labeled sets are available for training, existing data can be made coherent and coded for statistical purposes. The focus will mainly be on textual data describing the economic activity performed by the firm or the local unit, the correction and standardization of addresses and, to a lesser extent, the computation and control of tax numbers. Each technique can easily be implemented with the packages available in R[15].

By *text mining* (or *natural language processing*, NLP) is generally meant a set of techniques applying data mining concepts to textual data in order to extract quantitative information from qualitative and performing *information retrieval*, i.e. "finding material of an unstructured nature that satisfies an information need" (Manning et al., 2008). The object of the analysis is usually a *corpus*, that is a set of *documents*. A document in a database is a single observation: in our dataset, documents are textual data referred to a firm or local unit.

In order to be manipulated, the document is generally converted into a list of words. As well as in data mining, a common aspect of any text mining technique is that the analysis is preceded by a pre-processing phase in which the data are prepared, cleaned and standardized. *Tokenization* consists of splitting a raw character; in practice, sentences are divided according to the elements composing them (words, numbers, punctuation); moreover, for the purposes of the analysis some elements can be neglected reducing dimensionality without losing relevant information. This is the case regarding *stopwords* like "the", "and", "be" and so on. The relevance of words follows a negative power law (Zipf's law); even though stopwords are the most frequent terms in the corpus, they don't reveal anything interesting for the purpose of the research (Manning et al., 2008). The same holds for punctuation characters and numbers.

Another strategy for dimensionality reduction consists of replacing lists of words with their *synonyms*, so that the training set of available examples in a classification process is wider with less noise. For instance, both the words "tinsmith" and "plumber" can indicate the same profession, so

15 As far as text mining is concerned, for text preprocessing, tokenization, construction of document-term matrix and other tasks the package TM is available. NLP is also a rich package for natural language processing. Linear discriminant analysis (as well as other useful techniques) can be performed using MASS while the packages caret, randomForest and ParallelForest are available for growing trees and forests.

that only one word can be chosen thus reducing the number of columns (that are used as predictors) in the DTM; moreover, in reducing dimensionality it is likely that the two tradesmen are classified to perform the same economic activity even though their textual descriptions are formally different[16]. Words are also converted to their linguistic root eliminating affixes and retaining stems (from which it follows the term *stemming*); for instance, the past participle "gone" becomes "go". Semantically, terms like "unemployment" and "unemployed", "growth" and "growing", "increase" and "increasing" etc. are supposed to belong to the same group insofar as a common underlying topic is detected (in this case, the situation of the economy[17]) and they can be stemmed as "unempl", "grow" and "incr", respectively. Words in uppercase are converted to lowercase: the proper name "America" becomes "america" and the same happens for terms in uppercase at the beginning of a sentence. Pre-processing allows more matches between the words contained in the corpus and the ones listed in the predefined dictionary.

After preprocessing, texts can be algebraically represented in matrix form. Suppose $d \in \{1,2,\dots,D\}$ the set of documents; if each unique term in the corpus is indexed by $t \in \{1,2,\dots,T\}$, then $\boldsymbol{M}$ is a D x T *document- terms matrix* (DTM) with elements $m_{d,f}$ indicating the frequency of the t-th therm in the d-th document. Vectors of the DTM can also be simple binary indicators for the bare occurrence of a word in a document.

Preprocessed and pre-classified documents are converted into matrix form becoming the training set feeding the classification algorithm which will be used to assign unlabeled documents to groups. In text mining, classification is a task aimed at learning a classification function mapping documents to classes (Manning et al., 2008). Three classification algorithms will be suggested and exploited: linear discriminant analysis and random forest.

i) Linear discriminant analysis

LDA[18] is a linear method for classification. Suppose our output is represented by K classes and we want to divide the input space in regions labeled according to the classification itself; LDA is a technique aimed at identifying the best linear boundaries separating the input space. The idea is about assigning the observed data to the class k maximizing the class posterior:

16 It is superfluous to emphasize the language and domain specificity of textual preprocessing.

17 Topic detection is a typical unsupervised learning task.

18 LDA should not be confused with Latent Dirichlet Allocation, an unsupervised learning technique by Blei et al. (2003).

$$Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \propto f_k(x)\pi_k. \qquad (1)$$

From (1) follows that all we need is an assumed class density $f_k(x)$for k = 1,2, …, K and an estimation $\hat{\pi}_k$for the prior probability of the class k, π_k. Comparing how likely two classes are, we take the following ratio:

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=l|X=x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l). \qquad (2)$$

Intuitively, the higher the ratio, the more likely the observation is to belong to class k. Since LDA is based on the special assumption that all classes are distributed as gaussians with constant covariance matrix $\Sigma_k = \Sigma$, quadratic terms disappear and the decision boundary is linear on x. The boundaries separating observations belonging to class l from that ones belonging to k are the set of points such that $Pr(G = k|X = x) = Pr(G = l|X = x)$, from which follows the linear discriminant function (Hastie et al., 2009)[19] :

$$\delta_k(x) = \log\pi_k + x^T \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k. \qquad (3)$$

ii) Naïve Bayes

As for LDA, the Naïve Bayes classifier starts from the Bayesian concept of conditional probability. The probability that a document d belongs to the class k is

$$Pr(G = k|d) \propto \pi_k \prod_{1 \le j \le nd} Pr(t_j|k) \qquad (4)$$

where $t_j \in \{t_1, t_2, \ldots, t_{nd}\}$are nd tokens in d and $Pr(t_j|k)$is the probability of a given word to

[19] The estimations for the unknown parameters obtained from the sample are:

- $\hat{\pi} = N_k/N$
- $\hat{\mu}_k = \sum_{gi=k} \; x_i/N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{gi=k} (x_i - \hat{\mu}_k)\,(x_i - \hat{\mu}_k)^T/(N - K)$.

occur when the class is k (see Manning et al., 2008). As for LDA, the document is assigned to the class k such that the posterior probability (4) is maximized. What mainly distinguishes the NB classifier from the LDA is the assumed conditional independence of features, so that the right-hand side of (4) reduces to a joint product of conditional probabilities, i.e.

$$Pr(d|G = k) = Pr(t_{1,}t_{2,}\ldots, t_{nd}|k) = \prod_{1\leq k\leq nd} Pr\big(t_j|k\big). \qquad (5)$$

Furthermore, the probability of a token given the class is assumed to be independent on its position on the document (positional independence). Although those may appear strong assumptions for our purposes (terms in a document are unlikely to be independent), nevertheless they simplify computations.

Finally, in the Bernoullian version of the NB classifier, the bare presence of terms in the documents is considered -instead of the count of occurrences- using binary indicators.

iii) Random forests

Random forests were proposed by Breiman (2001) and combine decision trees and the concept of bootstrap aggregation underlying the bagging technique. The main idea is about averaging a set of (de-correlated) trees resulting from B different bootstrapped samples reducing variance[20]. The trees are estimated considering a random subset m of the p predictors[21] in order to stop the models from being too similar and, consequently, correlated. When the problem is continuous the prediction is represented by the average of the B estimated trees, whereas for qualitative variables the outcome is selected among B classes according to majority vote. See Biau and Scornet (2016) and Hastie et al. (2009) for further details on random forests.

Text mining for data cleaning

We apply text mining techniques and classification algorithms exposed above to the available

20 The underlying idea is rather trivial. In fact, such an approach relies on the fact that the variance of the mean sample (for *iid* elements) is σ^2/B so that, as the number of estimated models increases, the variance goes to zero. Trouble arises if the average models are similar themselves, i.e. there are similar combinations of predictors and cutpoints (j, s) in which the predictors' space is partitioned so that the obtained models are highly correlated.

21 For classification $m \approx \sqrt{p}$. See Hastie et al. (2009).

textual data in order to correct errors and replace missing information in two of the three variables involved in the firm's classification strategy: economic activity code and address.

- **Economic activity**

We have a training set represented by the NACE Rev. 2 codes for the activities at the firm and local unit level and textual descriptions of such activities. The official NACE Rev. 2 handbook with combinations of codes and textual descriptions (not necessarily coinciding with the ones delivered by the firms to the business registers) can be included in the training set too. Whatever is the training set, some duplication can be performed in order to balance the sample as far as few examples for some category persist (see Härdle et al., 2009). In machine learning jargon, the textual description is the input and the code is the output forming the training set for the algorithm aimed at classifying (missing) economic activity codes at the firm and local unit level. Each row of the DTM represents a description of the economic activity associated to a firm or a local unit, each column is an element of the list of single terms and the absolute frequencies of a term in a document fill the DTM. Preprocessing strategies previously exposed are clearly aimed at reducing the number of columns in the DTM in order to make estimations easier.

A typical feature of text mining is the domain specificity of the dictionary and of the list of stopwords and synonyms. To give an example, in the financial domain bears and bulls are not merely animals but rhetorical figures describing market trends or investors' attitudes. Moreover, the purpose of the classification algorithm (and the weights assigned to single words) matters. For instance, the lion's share of economic activity descriptions are likely to begin with “The main activity performed by the firm is...”; in that sentence there are 7 single words likely to appear in each document, so that their informative power proxies to zero. Terms like “activity” and “performed” would form non-sparse column vectors of the document-term matrix increasing dimensionality and implying higher computational costs with a negligible contribution to classification accuracy. For that reason, similar words should be included in the list of stopwords. Nevertheless, frequent words like “farming”, “restaurant” and “handcraft” have dramatic importance; regarding this, the stemming process and the creation of a list of synonyms are vital. As far as the stemming is concerned, mapping a group of words like “farming”, “farmer” and “farm” to the same stem, like “farm”, has algebraic and practical consequences. Mathematically, the DTM's dimension is lower. In practice, a set of firms containing the term “farm” in their description are likely to be classified to the same group and labeled with the same Ateco code. A similar argument holds for the individuation of synonyms. In a description the use of several synonyms is likely to occur and different firms performing the same activity are likely

to have similar textual descriptions. For instance, descriptions of firms operating in the NACE Rev. 2 sector 15.11 -"Tanning and dressing of leather; dressing and dyeing of fur"- probably contain terms like "tanning", "dyeing", "hides", "skins", "chamois", "leathers", "manufacture", "scraping", "shearing", "plucking" and other similar ones, resulting in groups of words highly correlated. If each row of the DTM represents a firm, and different firms operating in the same sector are described by similar groups of terms, proportional row vectors of the DTM imply the system to be redundant and the risk of collinearity arises with possible matrices' singularity. Now it should also be clear how synonyms can reduce misclassification rate. The extent to which rich lists of synonyms and stopwords can help to overcome possible computational issues depends on the specific dataset and requires some "craft work" by the researcher.

As far as the importance of a single term is concerned, if the researcher is comparing company names in order to identify the common entity leading the firms, words with low frequency are much more informative than highly frequent ones. Practically speaking, common terms like "restaurant" or "hotel" play a crucial role in classifying firms' economic activity, whereas specific terms indicating the company name -which are rare across the corpus- can be neglected; instead, the latter terms are crucial in detecting legal or economic linkages between two different firms.

The next Sections are devoted at comparing three classification algorithms assigning each document to only one class[22]; the dominant algorithm is successively applied to replace missing information and to a lesser extent to correct errors in our data.

Classification performances evaluation[23]

The corpus is made up of 30,317 labeled documents belonging to 886 classes[24]; each document refers to a unique local unit and contains the textual description of the economic activity performed there (the input variable) and the relative NACE code (the output variable). 75% of data was randomly assigned without replacement to the training set and the remaining 25% to the test set,

22 This is known as a *one-of* problem; the alternative approach is the *any-of* issue.

23 A deprecated version of such work was previously presented during seminars and conferences. Results reported in Table 3 were slightly different due to the application of word frequency weighting rather than "Tf-Idf".

24 The huge number of possible classes represents a criticism and it is due to the decision to apply literally Eurostat's prescription about comparing economic activities at the 4-digit level. As a matter of fact, the researcher can decide to compare firms at the 3 or 2-digit level of economic activity code; as a consequence, this would lead to a higher number of firms found to be connected. As far as text classification is concerned, a high number of classes implies the probability of each class occurring being low, resulting in a high misclassification rate. In fact, keeping fixed the dimension of the training set and reducing the number of possible classes would increase the number of instances for each class, improving classification performance.

leading to a training DTM of 24,543[25] rows and 7,580 terms. Dimensionality reduction was performed using sparsity measures resulting in 1,249 predictors (words) used to train the model. Moreover, lists of Italian stopwords and synonyms were produced bearing in mind the domain specificity of the problem, as discussed previously, with special attention to highly correlated words. Just for the LDA algorithm, pairs of words with a correlation coefficient higher than 0.05 and occurring less than 15 times in the corpus were detected and one out of two was removed from the DTM in order to prevent singularity issues. In addition, groups of collinear vectors were identified and eventually removed. As regards the NB classifier, the bare occurrence of a word in a document, as expressed by boolean indicators, was used in order to build the DTM. For the random forest, 10, 50, 100 and 300 trees were estimated with node size of 5. Results stabilize starting from 100 trees, consistently with the existing literature (see Hastie et al., 2009). Performances of the classification algorithms are reported in Table 3. As it is shown, RF outperforms NB and LDA algorithms. The relatively high misclassification rate in the test set[26] should be handled carefully and needs to be clarified. First, the fact that the percentage of misclassified observations falls from 46.7% to 29.9% when considering the classification at two-digit level means that the algorithm is sufficiently capable of detecting the economics activity at least at the division-level, that is, identifying approximately the goods and services produced, their destination and the technology involved in the production process. It should also be noted that training data contain labeled examples with two, three or four digits and the NACE Rev. 2 codes with three and four digits are clearly a subset of two-digit examples. Including in the training set only examples with, let's say, three or four digits would clearly reduce the number of classes and the consequent error rate; in order to maintain internal coherence, all the observations should be reclassified and converted to an established level of detail. The second caveat is the following: from manual checks we realized that it can be the case in which the class assigned to an observation by the algorithm is more accurate than the one reported in the dataset, highlighting the presence of registration errors in the data. As a result, the test error appears to be overestimated due to the presence of wrong examples in the training set itself. On the other hand, the low quality of the dataset is likely to be mitigated by the inclusion of the official NACE pairs description-code in the training data in order to combine official descriptions provided by NACE with the specific features of the dataset at hand[27], saving its peculiarities. Finally, after further checks we realized that some

25 To be sure that each class is involved in the training phase, we include 1,806 instances from the official NACE Rev. 2 list.

26 Theoretically, one of the advantages of RF is that it does not need a test set since softwares usually provide the OOB error rate. We applied the estimated model to a test set and actually OOB and test error coincide.

27 If it is the case, after some sample checks the researcher can replace wrong information deciding, as the extrema ratio, to re-label all the observations just relaying on the official NACE as a training set and getting read of the initial administrative information.

descriptions have no informative power and should be removed from the training set; nonetheless, when the data are big, those noisy observations can't be detected at hand. Bearing this in mind, we labeled 1099 textual descriptions (1000 of them were coded with 1, meaning "informative", and 99 were found to be uninformative and coded with 0) and we estimated a "preparatory" random forest identifying observations that are not useful for our purposes[28]. The resulting forest has an estimated OOB error rate equal to 0.45% and about 200 of the rough 30,000 initial documents were discarded from the sample accordingly, as they don't provide insights on the economic activity actually performed by the firm or local unit. Then, we re-estimated the random forest exploiting the new, slightly smaller and less noisy sample resulting from the previous step. As a result, we expect the random forest classifying the economic activity based on the textual description to be more accurate having reduced noise. Indeed, the random forest estimated using such a new and "cleaner" dataset is performs better as it can be shown in Table 3.bis.

Table 3: Classification performances

	Training set misclassification rate		**Test set misclassification rate**	
	2 digits level	**4 digits level**	**2 digits level**	**4 digits level**
Random Forest	18.6%	30.2%	29.9%	46.7%
LDA	35.6%	49.4%	40.7%	56.3%
Naïve Bayes	44%	60%	41%	58%

Notes: random Forest(formula = y ~ ., data = train sparse, ntree = 300, nodesize = 5); type of random forest: classification. Number of trees: 300. No. of variables tried at each split: 53. OOB error rate: 46.9%.

Table 3.bis: Classification performances (reduced training set)

	Training set misclassification rate		**Test set misclassification rate**	
	2 digits level	**4 digits level**	**2 digits level**	**4 digits level**
Random Forest	18.2%	29.5%	29.7%	45.9%

Notes: random Forest(formula = y ~ ., data = train sparse, ntree = 300, nodesize = 5); type of random forest: classification. Number of trees: 300. No. of variables tried at each split: 53. OOB error rate (at 4-digits level): 46.9%. More than 200 documents were removed from the training set as the textual descriptions reported were not informative about the economic activity.

The resulting random forest, which performances are reported in Table 3.bis, is finally used to impute missing NACE codes as far as textual descriptions are available. Two out-of-sample datasets composed by, respectively, 460 and 474 statistical units were generated and the missing economic activities were consistently imputed. A research assistant and the author performed, independently,

[28] The typical examples of uninformative descriptions are "secondary office" and "legal seat": that information refer to the organizational role of the local unit rather than the economic activity actually performed there.

manual checks on the NACE codes predicted by the random forest, comparing the textual description of the economic activity performed by the firm with the one reported by the official NACE manual and associated to the predicted code. In both samples the 89% of predicted codes resulted to be correct at least at the two-digit level; at the four-digit level, the 80% of observations was considered to be correctly classified in the greater sample while the percentage of correct classification was estimated to be 85% in the smaller one[29].

To finalize, those findings suggest that the test error in Table 3.bis is likely to be over-estimated, leaving room for the researcher to apply random forest to re-label the whole dataset combining administrative information and external ones in order to correct errors arising directly from the data source.

- **Localization**

Since our implementation of the classification algorithm is based on detecting exact matches between addresses of different firms, the standardization of formats plays a crucial role. Public administrations are rich data sources but bureaucracy, like technology, evolves over time, and qualitative standards for the collection and recording of information can vary. Consequently, data likely contain collection and registration errors or, alternatively, data can be correct but formats are less coherent. Three alternative approaches for the correction and standardization of addresses are proposed: the first approach is easy and reasonable effective and consists of the manual correction of more evident and frequent errors while the second one exploits the ggmap[30] R package and the last method is equivalent to the one applied for the economic activity code.

If a list of correct addresses is available, namely univocal combinations of roads and cities, wrong localizations can easily be detected through comparison with such a list. Addresses excluded from the official list of the municipality should be ranked and at least the most frequent errors should be corrected.

Another approach exploits ggmap and enables the automatization of the correction with R. There are two possible drawbacks for such an approach: the first one is due to the fact that Google allows a limited number of daily uses; the second possible concern regards coherence: it is possible that also addresses that were recorded correctly need to be adapted to the Google map format.

The third approach is based on ML and NLP and it is analog to the one applied for the

29 The referred material concerning out of sample manual controls is available on request at giacomo.caterni.91@gmail.com or giacomo.caterini@unitn.it .

30 D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

economic activity. Textual data are pairs of streets and towns that should be tokenized and preprocessed before NB, LDA or RF are applied; obviously, when dealing with addresses, synonyms and stopwords play a negligible role if compared with the preprocessing of typical textual data. In fact, a certain amount of manual work is required in order to prepare examples of unstandardized addresses and cities with their equivalent correct addresses. In that application, the output is represented by the correct address (possibly the official one if administrative lists are available) and the input is given by the combinations of roads and cities. A possible strategy exploits the fact that, after tokenization, remotion of stopwords (including initial letters for abbreviated names) and conversion to lowercase, the document-terms matrix of the wrong addresses appear quite similar to the one generated for the correct ones, so that one can proceed as follows: duplicate the column of correct addresses and use one of the two as an output vector in the training set; combine the other column vector, in which addresses are also correct, with the vector of cities and use those two column vectors as an input matrix to train the model; finally, use the model previously estimated to classify the wrong addresses. In building the training set, the relative importance of different locations should be taken into account and its composition should reflect, in some extent, the frequencies with which addresses effectively occurs in the dataset. Moreover, labelling a sample of wrong addresses by hand would improve classification performances.

As the quality of our data is fairly good, we preferred the first approach. The alternative approaches are still valuable solutions for the cases in which data are particularly noisy.

- **Tax number**

In this section we conclude the part dedicated to preprocessing and we briefly discuss issues that a researcher can face when dealing with individuals' ID or, more specifically, with their tax numbers. The easiest possible case is the one in which all the information needed to compute the tax number are available, but the desired code is not reported in the dataset, for some reason. In such a favorable case, an easy solution for Italian tax numbers is represented by the R package "ifctools". It can also be the case in which the researcher is uniquely interested in determining the gender of an individual; in R the gender can be obtained from the name -for a limited number of countries- using the package "gender". Considering that the task of detecting the sex from the name is extremely language-specific, an alternative approach exploiting text mining consists of building a training set, labelling names by hand and classifying missing observations using one of the classification algorithms previously exposed.

5 The Implemented Algorithm for Births

Once data are preprocessed, births and deaths are classified. The first step of the algorithm identifying new births consists of the definition of the reference period. In our implementation we compare firms[31] which were activated between *01/01/t* and *31/12/t*[32] for the first time with: a) formally different firms activated before *01/01/t* and still active at time *t*, b) firms ceased between *01/01/t-1* and *31/12/t* independently of their year of activation. It should be noted that firms activated in *t* are compared with other firms activated or ceased in *t* itself; in fact, it can be the case that a firm created in *t* exhibits continuity with another firm -still active or apparently dead- activated in the same year. The rationale is about detecting cases in which a firm formally ceases its activity in *t-1* or in *t* and is reactivated in *t* with a new registration number, as described in Section 2. From the way in which the period of analysis is defined it follows that the issue of reactivation of dormant firms is partially neglected. According to Eurostat (2010), dormant firms temporarily cease their activity and are resumed within 24 months retaining their old identity number (this is relevant for firms running seasonal activities, for instance). As the date of first activation is known -and new activations in our dataset receive a new identity number- they cannot be reactivations by definition. Moreover, considering a reference period spanning from *01/01/t-1* to *31/12/t* we theoretically face the extreme case in which a firm activated, say, on *31/12/t*, is matched with a firm ceased on *01/01/t-1*, i.e. exactly 24 months earlier. Based on our experience, demographic events mainly occur at the end of the year (or at the beginning, depending on the viewpoint) so that such a scenario is extreme but far less likely. Finally, the width of the interval of comparison affects the results, and consequently statistics on business demography may be under or overestimated, and the chosen interval depends on the available set of information.

For the matching process we focus on three kinds of information: a) demographic data about the firms, especially the addresses of all the local units, b) tax numbers[33] of legal and natural persons controlling[34] the firms and c) the 4-digits NACE code of economic activity. The full address[35] of firms

---

31 Legal forms such as cooperatives, consortiums and nonprofit organizations are excluded from the statistics on business demography. Holdings are excluded too (Ateco 642).

32 Firms created or ceased before *01/01/t-1* and after *31/12/t* are not relevant.

33 If tax numbers are not available, the full name and eventually the date of birth should be used to identify univocally individuals; for the legal entities controlling the firms, similarity measures on the company name should be elaborated.

34 The underlying idea is the following: two firms are linked if there are some key individuals or entities in common between them. We define two controlling positions: shareholder with a share of ownership greater than some threshold (in our implementation a controlling shareholder is considered to own at least the 30% of the company) and other individuals (non-owners) holding key positions. The choice of the threshold for the share of ownership and the definition of controlling positions are crucial for the final result. Roughly speaking, the more restrictive the definition of control, the less subjects are involved in the matching process and the fewer linkages are found, resulting in a higher number of real new firms. Controlling positions are reported in the appendix.

35 Addresses in which offices of professionals or startups incubators are located should be handled with care. Some manual checks

and local units is the proxy used for the localization; as a proxy of the legal entity, we use the tax numbers of individuals or entities involved in the firm at time *t,t-1*and *t-2*[36] while a 4-digit code describes the economic activity performed by the firm or the local unit. If more than one economic activity is conducted in the same location, we consider just the prevailing ones, where a ranking is indicated. All the variables involved need some preprocessing in order to be standardized and made consistent. In this regard, since the algorithm is based on detecting exact matches between strings referred to (formally) different firms, trivial cleaning is required, like removing blanks[37] and homogenizing the presence of upper and lowercases. Instead, addresses should be preprocessed more carefully, exploiting the methods described in Section 4. The proxies chosen for economic activity and legal control make the comparison particularly easy. Moreover, if textual data describing the economic activity are available, missing NACE codes can be replaced as suggested in Section 4. For this purpose, the researcher should decide whether the available data can be exploited to train and test the model or not. In fact, the researcher can classify the economic activity just for those firms for which descriptions (or some textual information about the activity can be retrieved from the company name) are available but the NACE code is missing, or she can decide to re-classify the economic activity performed by each firm and local unit exploiting textual descriptions. In the first case, the researcher can assess the accuracy of the classification recorded at the business register based on the misclassification rate on the training set: indeed, it may happen that, once the model has been trained, its classification is more accurate than the one supplied to the business register. Instead, if the second approach is chosen, the official NACE classification constitutes the training set, that is, pairs of textual descriptions and official codes: the inputs and outputs of the training data. In the latter case, statistical units are eventually assigned to new classes which don't necessarily coincide with the administrative one provided by the officers. To conclude, the chosen strategy depends on the perceived quality of the administrative data at hand.

For our implementation we arrange the data as follows:

– the statistical units are univocal IDs identifying each firm (the business register code is

need to be performed in order to exclude - to some extent at least - those localizations from the matching process. Our strategy consists of identifying those (full) addresses that are very frequent in the dataset and compare them with the official register of accountants of the province; several different firms can be formally located at the same address solely for practical purposes but they do not necessarily perform their economic activity there. As a consequence, if two firms perform, let us say, the same economic activity, and both of them are formally located at the same office, the matching process will misclassify one of them (i.e. the potential startup) detecting a mistaken link of continuity. In order to overcome such an issue, we exclude from our control combinations of localization and economic activity when the address coincides with a critical one. On the other hand, a match occurring over the combination of address and legal unit is a stronger hint for continuity.

36 Individuals can enter or exit from the firm at any time. Our data describe the situation at the end of each period and information about individuals or entities involved in the ownership or appointed at the boards can be overwritten. In order to trace individuals and entities we take data at the end of *t*, at the end of *t-1* and at the end of *t-2* so that we can observe the whole landscape from the beginning of *t-1* to the end of *t*.

37 For this purpose, the function *gsub* coupled with some regular expressions are particularly useful in R.

constant for any local unit belonging to a single firm);

– there are 3 variables, one for each measure of continuity used in the matching process, namely localization, economic activity code and legal unit;

– for each firm there are multiple observations; each observation consists of a local unit (with full address), the economic activity performed there and the ID of a controlling individual or entity;

– individuals and entities are duplicated, in the sense that the data frame includes each possible combination of localization/activity and legal unit.

For our purposes, any combination of localization and activity, activity and legal unit, legal unit and localization referred to a firm is compared pairwise with other (formally different) firms in order to detect duplicates. Roughly speaking, it is enough for us to show that there is a connection between a firm activated in *t* and another preexisting one for at least one local unit - it does not matter in which localization or for which controlling subject; if two out of three variables coincides for two different firms, this is a sufficient insight of continuity.[38]

The comparison exploits the function *dupsBetweenGroups*[39] which basically relies on the R function *duplicated,* which is applied too. The function *dupsBetweenGroups* returns TRUE if there are duplicated[40] observations between two different groups (in our case each group coincides with a firm). The function verifies whether two different statistical units have one or more variables in common at the same time. The matching process consists of three steps in which the three pairs of continuity measures are verified to coincide jointly for two different firms. Moreover, the function *duplicated* returns TRUE for the mere presence of duplicates, also within the same group, but has the advantage that the first observation of a sequence of duplicates is labeled FALSE: for our purposes this is the equivalent of saying that there is one and only one previously existing firm generating the others. By ordering data by date of activation, registration number and by the progressive number of the local unit[41], if available, and by using both the functions we are able to identify which firms are connected to a previously existing one and according to which criterion; moreover, the first firm in the sequence is the one which "generated" the others and is the only one which can eventually be considered as a new birth. The final result of the algorithm is merely obtained by excluding from the theoretical set of firms activated at time *t* the ones that were labeled (TRUE, TRUE) according to

38 For statistical purposes it does not matter if a formally new firm α that starts its activity β in the localization γ at time *t* is exactly the same firm which used to perform the same activity β in the localization γ at *t-1*; what matters is that from the statistical point of view, there is no creation of new production factors since a firm ceased and a firm started in a zero-sum game.

39 The function, realized by the R contributor W. Chang, is available online at: http://www.cookbook-r.com/Manipulating_data/Comparing_data_frames/ and the R syntax is reported in the appendix.

40 Missing values and blanks should be removed in order to avoid mistaken matches.

41 This is like saying that the first main office of the firm comes first in our data.

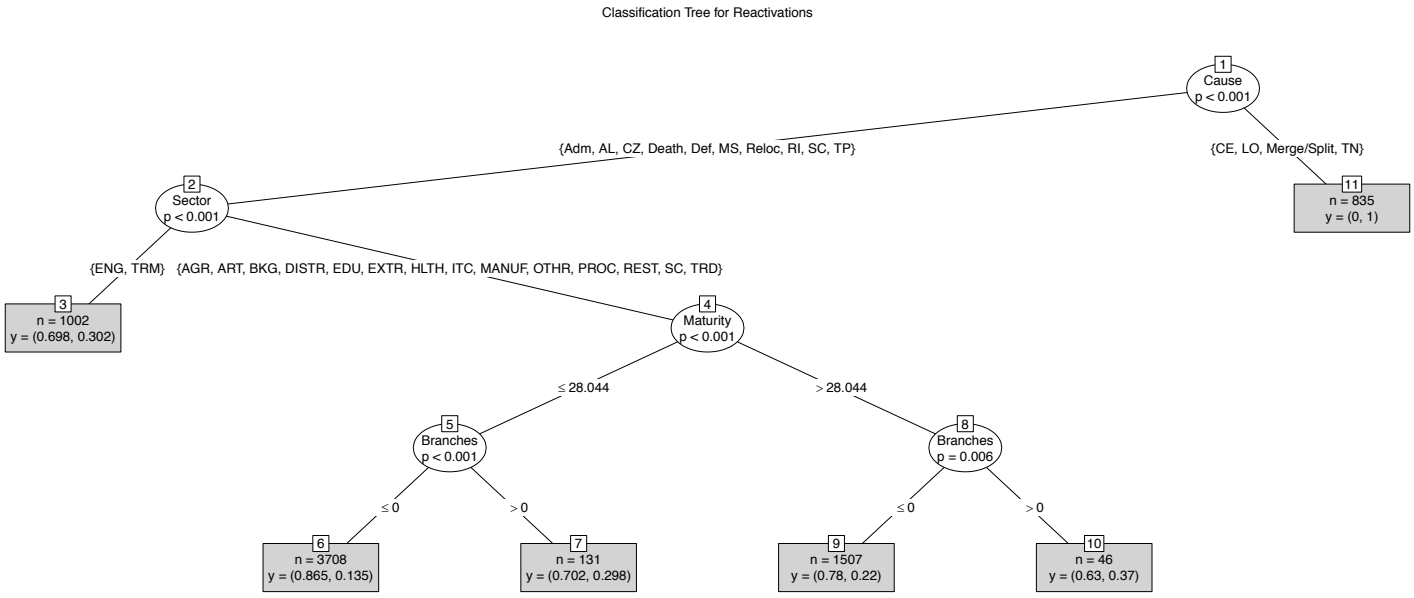
both functions.

6 Forecasting Reactivations with Random Forest

We identified ceased firms and reactivated ones between 2012 and 2017. We applied a matching process analog to the one described for births, generating a labeled set of 15,836 examples used to feed the classification tree and the random forest predicting whether a ceased firm will be reactivated or not. As input variables we used categorical and quantitative information about the legal form of the firm, the NACE Rev. 2 code for the activity, the cause of cessation, the municipality, the gender of the controlling person and their age, the average age of individuals controlling the firm and the standard deviation, a proxy of the firm's maturity represented by the number of years between activation and cessation, a measure of concentration of ownership represented by the percentage of shares held by the largest shareholder[42], the number of branches the firm had, the year and the month of death of the firm capturing seasonality and some real-time information such as the spread between Italian BTP and German Bund, the unemployment and the market interest rate. In the spirit of nowcasting, if the economic cycle matters, real-time information on economic variables are included in the model in order to assess, at least to some extent, the magnitude of the stance of the economy on predicting deaths and reactivations at the firm level. As accounting data are available satisfactorily just for a subset of observations, namely 1,218 firms, we estimated two separate models with different sets of regressors: a conditional tree was feed with demographic predictors available for the whole dataset, while a random forest was estimated including demographic and accounting information for a smaller sample.

We divided the sample into a training set (80% of observations) and a test set (the remaining 20%). As the results obtained growing a forest are intrinsically more obscure, we first estimated a single conditional tree in order to gain some insights about the splitting process. The resulting pruned tree is reported below: the algorithm predicts a binary outcome in which the apparently dead firm takes value one if it is reactivated and zero otherwise. Values contained in the brackets inside the gray rectangles refers to the probability of each outcome to occur.

42 See Gul et al. (2010).

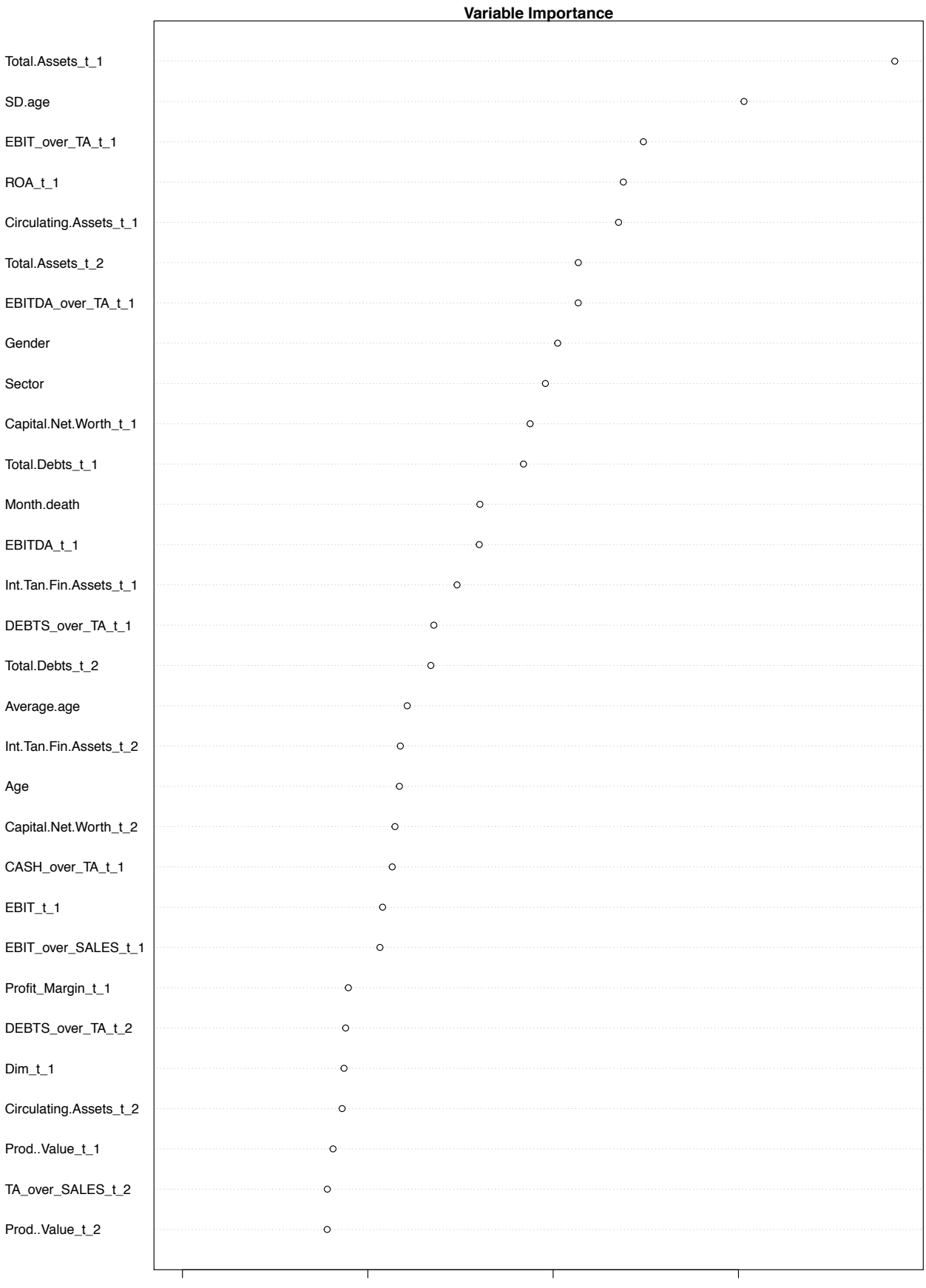


The first split occurs according to the cause of cessation. Not surprisingly, cessations due to mergers and splits are always predictive of a continuity linkage. The branch on the right-hand side includes cessations due to death of the owner and default of the firm. Then the activity performed matters, the age of the controlling individual and the number of local units. The classification error for the test set is about 17%.

A drawback of using trees is represented by the poor out-of-sample performances and high variability despite the good in-sample accuracy (see Hastie et al., 2009 and Breiman, 2001); being aware of this, we estimated a random forest with 500 trees including accounting data among predictors, although the number of observations (1,218 firms) is relatively small. In such a model we didn't include the cause of death, as in some cases (such as merges and splits) a given cause of death implies by definition a reactivation.[43] A measure of the predictive importance of variables based on Mean decrease Gini and the confusion matrix of the random forest are reported below.

The reported Mean decrease Gini is a measure of importance of the regressors, which are ranked according to their predictive power. Not surprisingly, the lagged levels of total assets one and two years before the cessation, lagged values of EBIT, EBITDA and ROA as well as the total level of debts in the year preceding the cessation are the variables showing the greater predictive power. Among demographic data, the economic sector and the month of death (both capturing, in some extent, seasonality), are highly predictive, as well as the measures related to the age of people involved in the firm.

43 In order to reduce the dimensionality, we estimated a random forest including each of the 111 initial regressors; then, according to the resulting mean decrease Gini variable importance, we removed the ones with negligible predictive power and we estimated a new random forest exploiting 92 input variables.


Variable Importance
Total.Assets_t_1
SD.age
EBIT_over_TA_t_1
ROA_t_1
Circulating.Assets_t_1
Total.Assets_t_2
EBITDA_over_TA_t_1
Gender
Sector
Capital.Net.Worth_t_1
Total.Debts_t_1
Month.death
EBITDA_t_1
Int.Tan.Fin.Assets_t_1
DEBTS_over_TA_t_1
Total.Debts_t_2
Average.age
Int.Tan.Fin.Assets_t_2
Age
Capital.Net.Worth_t_2
CASH_over_TA_t_1
EBIT_t_1
EBIT_over_SALES_t_1
Profit_Margin_t_1
DEBTS_over_TA_t_2
Dim_t_1
Circulating.Assets_t_2
Prod..Value_t_1
TA_over_SALES_t_2
Prod..Value_t_2

Dealing with random forests, the OOB estimated error is equivalent to a test error. Nevertheless, we decided to split the dataset and to assess the goodness of the model testing it on a separate sample. In both sets, the great majority of the observations (around 55%) belong to the class "zero". The classification performances of the random forest are reported on Table 4.: as already mentioned, zero and one are the outcomes associated with the death and the re-activation of the formally ceased firm, respectively. The overall error rate in the test set is equal to 22.5%, while the estimated OOB error automatically obtained with the random forest was about 21%. Results should be interpreted as follows: in the test set, 134 observations (the 55% of the set) belong to the class labelled with zero (that is, those firms which are actually dead and will not be reactivated within two years) while the remaining 110 are labelled with one (namely, those firms which don't represent real deaths).

Table 4: Confusion matrix-random forest

		Predicted Outcomes		
		0	1	RECALL
Observed Outcomes	0	113	21	84.3%
	1	34	76	69%
	PRECISION	76.8%	78.3%	
	Overall Classification Error on the Test Set			**22.5%**

Classification performances are better for the class labelled with zero (15.6% misclassification rate) than for the class labelled with one: this probably reflects the composition of the dataset, which is slightly unbalanced. Nevertheless, the performances of the model in predicting whether an apparently dead firm will be reactivated or not are satisfactory enough considering that if the totality of observations were simply assigned to the class which is considered more likely based on the dataset at hand (that is, the class labelled with zero), the misclassification rate in the test set would have been greater than 40%: considerably worse than the observed result.

We can conclude that the model trained according to past data on deaths and reactivations performs satisfactorily in predicting whether a firm will reactivate itself after cessation. Nevertheless, there is room for further improvements and developments conditional on the availability of better and more complete accounting data to be combined to the demographic ones.

7 Results and Future Research

We implemented an automated version of Eurostat's algorithm distinguishing true startups from pre-existing firms and identifying entities which are reactivated after the cessation. We exploited machine learning, natural language processing and econometric tools for preprocessing and analyzing

granular data coming from administrative sources. In particular, missing data were imputed using random forests to classify textual information. Moreover, the application of random forest was proposed to predict whether an apparently ceased firm will reactivate in order to overcome the physiological time lag faced when dealing with business demography. As a result, in almost the 80% of cases we were able to classify firms correctly with room for further improvements. In fact, the sample considered to train the "full" model (namely, the one including both demographic and accounting data, in addition to information referred to the stance of the economy) is relatively "short" if compared to the dimension of the full dataset in which, on the opposite, other features regarding firms and people involved in them are extremely detailed. Moreover, those missing observations are unlikely to be distributed at random, being related to the dimension of the firms and its legal form. Furthermore, duplications could be performed in order to better balancing the sample and further ML tools (such as ANN) could be applied to classify the sector and to predict the reactivation of an apparently dead firm. Nonetheless, scholars, practitioners ad public officers can exploit the results of the present work, conditional to the availability of data, producing official statistics and assessing the quality of their dataset at hand. In fact, we have proposed a strategy to identify the economic activity performed by the firm using NLP and to verify in which extent administrative activations and cessations represent real firms' births and deaths, respectively. In this sense, distinguishing real start-ups from regenerated firms is of critical importance for academics when studying firms' performances and for policy-makers when the impact of a policy is assessed or a treatment (more specifically, one or more measures devoted to stimulate entrepreneurship) is assigned. The same rationale holds predicting firms' rebirth. Future research aimed at improving predictive performances will consist of the inclusion of further information in the space of predictors, i.e. financial data capturing the firm's specificity and real-time data on economic variables proxying the cyclical component in real time. Finally, in addition to forecasting reactivations, we aim to apply machine learning in order to assess how likely a startup is to survive.

References

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32

Castle, J. L., Fawcett, N. W., & Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, 210(1), 71-89.

Diebold, F. X. (2003, January). 'Big Data'Dynamic factor models for macroeconomic measurement and forecasting. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,"(edited by M. Dewatripont, LP Hansen and S. Turnovsky)*(pp. 115-122).

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*,*346*(6210), 1243089.

European Commission No 2700/98 of 17 December 1998 concerning the definitions of characteristics for structural business statistics.

Eurostat – OECD (2007), Manual on Businesses Demography Statistics.

Eurostat (2010), Businesses Registers. Recommendations Manual.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, *33*(1), 101-122.

Härdle, W., Lee, Y. J., Schäfer, D., & Yeh, Y. R. (2009). Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. *Journal of Forecasting*, *28*(6), 512-534.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer, New York, NY.

Gul, F. A., Kim, J. B., & Qiu, A. A. (2010). *Ownership concentration, foreign shareholding, audit quality, and stock price synchronicity: Evidence from China.* Journal of Financial Economics, 95(3), 425-442.

Ikudo, A., Lane, J., Staudt, J., & Weinberg, B. (2018). *Occupational Classifications: A Machine Learning Approach* (No. w24951). National Bureau of Economic Research.

Josefy, M. A., Harrison, J. S., Sirmon, D. G., & Carnes, C. (2017). Living and dying: Synthesizing the literature on firm survival and failure across stages of development. Academy of Management Annals, 11(2), 770-799.

Lev, B., & Gu, F. (2016). *The end of accounting and the path forward for investors and managers*. John Wiley & Sons.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge

University Press.

Roelands, M., van Delden, A., & Windmeijer, D. (2017). Classifying businesses by economic activity using web-based text mining.

Schintler, L. A., & Kulkarni, R. (2014). Big data for policy analysis: The good, the bad, and the ugly. *Review of Policy Research*, 31(4), 343-348.

Stough, R., & McBride, D. (2014). Big data and US public policy. *Review of Policy Research*, 31(4), 339-342.

Van Stel, A., Carree, M., & Thurik, R. (2005). The effect of entrepreneurial activity on national economic growth. *Small business economics*, *24*(3), 311-321.

Appendix

Table A1: Control positions	
"ACCOMANDATARIO DI SAPA"	Unlimited partner of limited joint stock company
"AFFITTUARIO O CONDUTTORE"	Lessee
"AMMINISTRATORE UNICO E PREPOSTO"	Sole director
"COEREDE"	Joint heir
"COLTIVATORE DIRETTO"	Direct farmer
"CONIUGE"	Consort
"EX SOCIO DI SOCIETA' DI PERSONE"	Former shareholder of partnership
"GESTORE DELL' ESERCIZIO"	Managing agent
"LEGALE RAPPRESENTANTE"	Representative
"LEGALE RAPPRESENTANTE / FIRMATARIO"	Representatives and signatories
"LEGALE RAPPRESENTANTE DI INCAPACE"	Legal representative of incapable
"LEGALE RAPPRESENTANTE DI SOCIETA'"	Representative of the company
"LEGALE RAPPRESENTANTE E RESPONSABILE TECNICO"	Legal representative and technical manager
"LEGALE RAPPRESENTANTE INTESTATARIO DEL TESSERINO"	Legal representative holder of the requirements
"PIU' AMMINISTRATORI"	Multiple directors
"PROPRIETARIO"	Owner
"PROPRIETARIO AUTORIZZATO A RISCUOTERE E QUIETANZARE"	Owner and credits collector
"RAPPRESENTANTE LEGALE DELLE SEDI SECONDARIE"	Representative of subsidiaries
"SOCIO"	Shareholder
"SOCIO ACCOMANDANTE"	Limited partner
"SOCIO ACCOMANDATARIO"	Unlimited partner
"SOCIO ACCOMANDATARIO D'OPERA"	Unlimited partner and employee
"SOCIO ACCOMANDATARIO E PREPOSTO"	Unlimited partner
"SOCIO ACCOMANDATARIO E RAPPRESENTANTE LEGALE"	Unlimited partner and representative
"SOCIO AMMINISTRATORE"	Shareholder and director
"SOCIO COMPROPRIETARIO"	Co-owner
"SOCIO CON FIRMA CONGIUNTA"	Shareholder and co-signatory
"SOCIO CONTITOLARE"	Co-controlling shareholder
"SOCIO DELL'IMPRESA ARTIGIANA"	Partner in craft business
"SOCIO DI SOCIETA' DI FATTO"	Shareholder of de facto corporation
"SOCIO DI SOCIETA' DI PERSONE RAPPRES."	Unlimited liability partner
"SOCIO DI SOCIETA' IN NOME COLLETTIVO"	Unlimited liability partner
"SOCIO E LEGALE RAPPRESENTANTE"	Shareholder and representative
"SOCIO E PREPOSTO"	Shareholder and technical manager
"SOCIO QUALIFICATO"	Partner holder of the requirements
"SOCIO UNICO"	Sole partner
"TITOLARE"	Principal
"TITOLARE DELL'IMPRESA ARTIGIANA"	Proprietor and craftsman
"TITOLARE DI LICENZA P.S."	Licensee
"TITOLARE E RESPONSABILE TECNICO"	Principal and technical manager
"TITOLARE FIRMATARIO"	Proprietor and signatory
"TITOLARE MARCHIO IDENTIFICATIVO"	Trade mark holder
"USUFRUTTUARIO"	Life renter
Notes: Table of controls positions	

Table A2: The function dupsBetweenGroups

```
dupsBetweenGroups <- function (df, idcol) {
# df: the data frame
# idcol: the column which identifies the group each row belongs to
datacols <- setdiff(names(df), idcol)
sortorder <- do.call(order, df)
df <- df[sortorder,]
# Find duplicates within each id group (first copy not marked)
dupWithin <- duplicated(df)
dupBetween = rep(NA, nrow(df))
dupBetween[!dupWithin] <- duplicated(df[!dupWithin,datacols])
dupBetween[!dupWithin] <- duplicated(df[!dupWithin,datacols], fromLast=TRUE) | dupBetween[!dupWithin]
# ============== Replace NA's with previous non-NA value ==============
goodIdx <- !is.na(dupBetween)
goodVals <- c(NA, dupBetween[goodIdx])
fillIdx <- cumsum(goodIdx)+1
dupBetween <- goodVals[fillIdx]
dupBetween[sortorder] <- dupBetween
return(dupBetween)
}
```

Predicting Start-ups Survivals with Machine Learning

ABSTRACT

In this paper we implement an automated version of Eurostat's algorithm aimed at distinguishing true startup endeavors from the resurrection of pre-existing but apparently defunct firms. We estimate a random forest in order to predict firms' performances and survivals probabilities and to assess whether regenerated firms are more likely to survive, if compared with real births, due to their previous experience.

Keywords: Business Demography; Classification; Text Mining.

JEL classification: C01, C52, C53, C55, C80, G33, L11, L25, L26, M13, R11.

1 Introduction

The study of the antecedents of firms' survival have been a central topic within the management literature; researchers, in particular, have massively investigated the reasons why firms survive according to different theoretical lenses (e.g., human and social capital theories, cognitive and behavioural theories, organizational ecology and evolutionary theories, etc.) with the aim of identifying under which conditions a firm is more likely to survive or cease (e.g., Aldrich, 1999; Agarwal *et al*., 2002; Abatecola *et al*., 2012; 2013; Amburgey *et al*., 1993; Bercovitz & Mitchell, 2007; Boeker, 1988; Bruderl & Schussler, 1990; Sutton, 1987). This attention, moreover, is increasingly growing because of the current economic situation in which firms are failing at faster rates than before (Govindarajan & Srivastava, 2016).

However, despite this great interest and the massive number of conceptual and empirical contributions that have been published (Bermiss & Murmann, 2015; Eggers and Song, 2015; Geroski *et al*., 2010; Hsu *et al*., 2015; Jenkins *et* al., 2014; Justo *et al*., 2015; Kalnins & Williams, 2014; Wilson *et al*., 2013), according to the review article by Josefy *et al*. (2017), some relevant grey areas should be still investigated. First, the authors suggest that studies about firms' survival and failure should be increasingly concerned in empirically assessing the contribution of variables positioned at different interacting levels, thus the individual, firm, industry and macroeconomic levels. Implementing a multi-level logic allows scholars to reach greater results in determining the significance of different sets of variables. Second, they raised the problem that antecedents of firms' survival can assume a different degree or have a curvilinear effect respect to the same factors when looking at firms' failure. For example, Zheng *et al*. (2015) and Su et al. (2011) respectively found that political ties and entrepreneurial orientation have a positive effect on long-term performance and a negative one on short-term survival.

Moreover, according to some other scholars interested in explaining firms' failure (Mantere *et al*., 2013; Shepherd *et al.*, 2015; Walsh and Cunningham, 2016), prior empirical papers did not take into account the benefits – in terms of imprinting features (Cope, 2011; Yamakawa *et al*., 2015) – of prior business experiences in the formation of new firms; these are here called *regenerated business entities*. Most of the studies, indeed, did not considerate the connection between ceased entities and the subsequent new-born firms, discarding the potential effects of their inclusion into the conducted analysis.

On this premise, the aim of this contribution is to start filling the prior highlighted gaps. In particular, in this study we have taken into account the antecedents of firms' survival and cessation collocated at distinct levels that, assuming different degrees, have differently affected the outcomes of new-born and regenerated business entities.

From that, the theoretical background of this work is mainly based on the famous Stinchcombe's (1965) liability of newness concept, which predicts that firms' failure has a higher rate in the first years of the organizations' lifecycle because of the lack of experience, trust among founders, and other variables that undermine building successful routines (Abatecola *et al*., 2012; Cafferata *et al*., 2009; Nelson and Winter, 1982; Hodgson and Knudsen, 2004; Abatecola and Uli, 2016). This has been enriched by the results of the empirical analysis that have been carried out over time on the variables that differently affect firms' survival and failure (e.g., Wholey *et al*., 1992; Fackler *et al.*, 2013).

As to accomplish the aim of this work, it has been implemented a statistical algorithm from Caterini (2018) aimed at classifying new firms recorded at the businesses register in order to distinguish effective births from regenerated ones. Machine learning techniques are implemented to identify the differences among these entities among the multi-level selected variables. Those techniques have been applied to a sample of more than 200.000 Italian firms whose financial and demographic data have been retrieved from the AIDA Bureau Van Dijk database and from the Chamber of Commerce of Bolzano, respectively, obtaining a sample of 12,023 new activations. Those firms were compared with active and dead previously existing firms in order distinguish start-ups from regenerated entities.

This paper heavily contributes to the research on firms' failure and survival, offering new strong insights – due to the sample size, the use of different statistical analysis techniques, and the identification and comparison of real new-born, ceased and regenerated firms – on the antecedents that can determine the exclusion of the firm from its competitive environment, its regeneration or its continuity. In this vein, results of this research will benefit scholars that want to understand under which conditions an antecedent can be positive or negative for firms' survival or failure; it stimulates new research on apparently failed firms and the study of antecedents of firms' failure and survival according to a multi-level logic through the exploitation of a tall and, most of all, large dataset[1] through machine learning techniques. New and established entrepreneurs, thanks to the results of this study, can better understand which individual, firm and environmental variables increase or not their chances of firm's continuity. Yet, this contribution is even more important if looking at the fact that empirical works investigating the liability of newness have mainly used US data (e.g. Freeman et al., 1983; Henderson, 1999), limiting the generelizability of the implications for the European firms (Cafferata et al., 2009; Abatecola et al., 2012). Finally, applying machine learning tools to improve predictions on firms' performances, there is room for supporting policy makers and start-ups' incubators efforts in targeting potentially weak firms, to accompany them at the beginning of their

[1] Dealing with machine learning, a dataset is said to be tall when there are many observations and large when there are many regressors.

life cycle and in designing appropriate policies and measures, once they are assessed to be effectively unexperienced rather than regenerated. Nevertheless, guiding shareholders in distinguishing potential successful companies from unsuccessful projects can help reducing information asymmetries and preventing the resultant rationing phenomena (Stiglitz and Weiss, 1981).

The reminder of the work is the following. First, the theoretical background on life and death of business entities is offered to the readers. Second, data description and data analysis, with an important focus on the machine learning technique, is detailed. Third, results of the estimated predictive model are presented. Discussion and conclusion end the contribution highlighting its limitations and the insights for future research.

2 Liability of newness and chances of survival

Some recent review works (i.e., Cafferata et al., 2009; Abatecola et al., 2012; Bakkery and Josefy, 2017) shown how the Arthur Stinchcombe's liability of newness concept has occupied an important niche in the management literature on organizational evolution over the last 50 years. In particular, Stinchcombe (1965) started from a seminal explanation of the "struggle for survival" (Darwin, 1859) – i.e., the continuing fight of biological organisms for their own existence – between new-born and older organizations, as to introduce the "liability of newness" construct, which explains why business entities are more likely to leave the competition within the first years of their life. In particular, he explained that the high mortality rate of firms mainly depends on the lack of experience of new-born firms and the presence of not familiar people:

> New organizations, especially new types of organizations generally involve new roles, which have to be learned; [. . .] The process of inventing new roles, the determination of their mutual relations and of structuring the field of rewards and sanctions so as to get the maximum performance, have high costs in time, worry, conflict, and temporary inefficiency (p. 148).

From that, the following scholars understood that the learning curve experienced by the firm during its evolution increases due to the development of some countervailing mechanisms of the liability of newness, such as routines (Nelson and Winter, 1982; Hodgson and Knudsen, 2004; Abatecola and Uli, 2016). The more processes are executed, the greater the experience gained and the chances to survive.

According to this first investigation, scholars became increasingly interested in theoretically and empirically explaining the internal and external factors that can overcome the failure of business entities at macro-, meso-, and micro- level of analysis. If these factors (e.g., lack of: trust among

founders, goodwill, strategic alliances, etc.) are promptly considered, firms can countervail them and augment their probability of continuity (Michael and Sung, 2005). Implementing them forms a "buffer" that can prevent the failure of the firm or, at least, to postpone it for few years or months, (this period is also called "honeymoon"; see, Fichman and Levinthal, 1991).

Empirical works on the study of firms' survival in the first years of life have differently posed attention on the following levels of analysis: individual, firm, sector, macroeconomic conditions.

Following this rationale, some scholars (Debrill et al., 2009; Yang and Danes, 2013; Revilla et al., 2016) found that businesses in which participate family members are more likely to survive the liability of newness, because of the prior social ties that facilitate internal communication. Moreover, firms that have more than a single founder survive longer than those started by individuals (Schutjens and Wever, 2000; Arribas and Vila, 2007).

H1a: *firms that involve family members have more chances to survive in the first years of life then firms who do not involve them.*

H1b: *firms founded by two or more individuals are more likely to survive than the ones with a single founder*

With reference to the firm-level of analysis, Freeman *et al*. (1983) – then supported by other subsequent empirical studies (e.g., Wholey *et al*., 1992; Varum and Rocha, 2012) – found that increasing the firm's size can somehow reduce the effects of the liability of newness. From that, it emerges that the liability of newness concept can be, in some terms, translated in a liability of smallness. Indeed, Aldrich and Auster (1986) formally formulates the new concept of the "liability of smallness", which explains that small firms, because of the lack of financial resources, the impossibility to attract skilled employees, and the difficulty to handle high fixed and administrative costs due for the governmental regulations, are more likely to die. This is also confirmed by the study that links firms' size and firms' age (e.g. Fackler et al., 2013). However, this connection has not been found in some countries, such as Italy (Audretsch *et al*., 1999).

Looking at the firms' performance, Aspelund et al. (2005) and Esteve-Perez and Manez-Castillejo (2008), found that new ventures' survival – respectively, within the Scandinavian technology and Spanish manufacturing industries, increased with the research and development (R&D) expenditures; however, this data is not confirmed in the Italian context where process or product innovations seem not to bring to a higher survival rate (Giovannetti *et al*., 2011). Other scholars found that the financial resources, in terms of capitalization, affect firm's viability (Caves, 1998; Bales, 2005). About the

legal form, Mata and Portugal (2002) find that unlimited liability firms are more likely to fail than limited liability companies, while Esteve-Pérez and Llopis (2004) found opposite results.

H2a: *Italian big size firms have the same chances to survive in the first years of life then Italian small-medium firms.*

H2b: *Italian firms with greater investments in R&D have the same chances to survive in the first years of life than firms who invest less.*

H2c: *firms with greater capitalization have more chances to survive in the first years of life than firms with less capitalization.*

H2d: *firm's legal form affects company survival*

Beside these internal causes of failure, the management literature has found over time a high load of external factors, in terms of macroeconomic and sectorial dynamics. Some major streams in the research related to the sectorial dynamics and firms' survival are about industry location, market growth and industry innovation. In this vein, despite Porter (1990), empirically supported by Delgado et al. (2010) and Wennberg and Lindqvist (2010), suggested that knowledge spillovers in geographically concentrated industries stimulate growth, Sorenson and Audia (2000) found that geographic concentration contributes to firm failure; in other words, higher agglomeration is associated with higher firm mortality rate (see also Folta et al., 2006).

About market growth, some scholars (Mata et al., 1995; 1999; Mata and Portugal, 1994; Strootman, 2006; Campbell et al., 2012) found that market growth is associated with firms' survival rate; however, if there is a high number of entries in the market, the chances of survival are reduced (Mata and Portugal, 2002; Segarra and Callejón, 2002). About industry innovation, if firms enter in an innovative market, their chances of survival are few, as supported by several studies (Audretsch, 1995; Cader and Leatherman, 2011; Ejermo and Xiao, 2014).

H3a: *firms' geographic concentration affects the chances of survival*

H3b: *market growth increases the chances of firms' survival*

H3c: *the increase of market entries diminishes the chances of firms' survival*

H3d: *firms that enter markets featured by high technology rates have less chances to survive*

Looking at the macroeconomic trends, Everett and Watson (1998) found that small businesses have higher failure rates during periods in which there are high unemployment and interest rates. Fotopoulos and Louri (2000) also confirmed that mortality rates decrease with positive economic cycle; vice-versa, Box (2008) found that firms formed in periods of positive macroeconomic conditions have higher survival rates. Moreover, studies on the labour cost demonstrated that if the cost of a workforce unit, thus the real wage (money wage/consumer price index x 100), is high, it has a positive relationship with firms' failure in the first years (Platt and Platt, 1994; Salman et al., 2011).

H4a: *firms have less chances to survive when the unemployment rate is high*

H4b: *firms have less chances to survive when interest rates are high*

H4c: *firms have less chances to survive when the real wage is high*

However, as highlighted by recent works (Mantere *et al*. 2013; Shepherd *et al*., 2015; Walsh and Cunningham, 2016), it has not been considered over time the beneficial imprinting features released by prior ventures to entrepreneurs, who can take advantage from them in creating new firms (Cope, 2011; Yamakawa *et al*., 2015). In this work, these new business entities are called as *regenerated* (or *apparently died*) firms, thus the ceased business entities that find a new life in organizations with a new ID number released by the business register, but that operate in the same sector, geographical area or under the control of the same ownership. The buffer of founders that previously failed, in terms of gained experience, has a positive effect on new-born firms (Dias et al., 2017; Amankwah-Amoah et al., 2016); the committed mistakes are less probable to appear. In this vein, Dyke *et al*. (1992) and Renski (2015) found that the prior experience of the founder/s in the industry has a positive effect on firm's survival.

H5: *firms owned by founders with prior entrepreneurial experience have more chances to survive*

In order to test the validity of the hypothesis stated above and to assess the relevance of the introduced features for the survival of start-ups, a mix of qualitative and quantitative regressors at the firm level as well as the sectorial and macroeconomic one will be built up. Those regressors will constitute the

inputs to feed tree-based algorithms having as possible outcomes the firm's survival, firm's death and its reactivation after apparent death.

3 Methodology: Trees and Random Forests

Classification problems are typical tasks in machine learning, whom algorithms reveal to be extremely powerful especially when multiple classes are dealt with.

Trees (Morgan and Sonquist, 1963, Breiman et al., 1984) are techniques capable to handle nonlinear relations between inputs and outputs. The algorithm partitions the set of inputs in sub-regions and a model is fitted in any of them. At any step, a variable and a split-point are chosen to separate the feature space: in regression problems the model is fitted predicting the response variable as the average of the observed y_i in any region and the split-point is set to reach the best fit; then, other split-points are chosen for the previously obtained regions and other variables are split until a stopping rule is reached. Consider the case of N (x_i, y_i) pairs, with $x_i = (x_{i1},...x_{ip})$, that is p inputs and N observations, and suppose to divide the space of regressors in M regions $R_1, \dots R_M$ and represent the constant as c_m (Hastie et al., 2009), then:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{1}$$

from which it clearly follows that, if $f(x)$ is chosen minimizing $\sum(y_i - f(x_i))^2$, the best $\hat{c}_m$ is the mean of y_i conditional to the region R_m to which the indicator function refers. Practically, starting from two regions R_1 and R_2, partitioning consists in finding the pairs (j, s), *ie* the predictor j and the split-point s, minimizing:

$$\sum_{x_i \in R_1(j,s)} (y_i - \hat{y}_{R1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{y}_{R2})^2, \tag{2}$$

and repeating the process for each predictor and each cut-point (see James et al., 2013). As for polynomial regression, we can increase the accuracy of the fit as much as we want simply increasing the number of regions in which the features are partitioned and, consequently, the number of terminal nodes. When the phenomenon under analysis is likely to be nonlinear, trees perform better than linear regressions or logistic models: even though regression trees work linearly via the sample mean, partitioning the input space the researcher is able to capture nonlinear relations. The extent in which the feature space is partitioned affects the bias-variance trade off since the size of the tree (namely, the number of branches and nodes) governs the complexity. Operationally, a large tree T_0 is estimated

and then it is gradually *pruned* (that is, internal nodes are removed), reducing complexity; at that stage, cross validation can be applied in order to select the tuning parameter based on the error rate in the test set. When the tuning parameter equals zero the obtained subtree coincides with T_0. Alternatively, pre-pruned conditional trees can be grown. See Hastie et al. (2009) and Varian (2014) for further details and economic applications.

As the output of the supervised problem is categorical, classification trees work analogously to regression ones: the constant outcome predicted for a given partition of the predictors space of the test set is represented by the most frequent class in the same region of the training data and, estimating the tree, squared errors are replaced by entropy measures and the goodness of the tree is evaluated according to misclassification rates.

A drawback of using trees is represented by the poor out-of-sample performances and high variance despite the good in-sample accuracy (see Hastie et al., 2009 and Breiman, 2001). Random forests were proposed by Breiman (2001), combining decision trees and the concept of Bootstrap aggregation underlying the bagging technique. Bagging is a variation on a theme of the Bootstrap. It consists in averaging predictions obtained based on *B* bootstrap samples in order to reduce the variability of learning methods. In formulas (Hastie et al., 2009 and Breiman, 1996):

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x) \tag{3}$$

increasing the accuracy of the model as $B \to \infty$ and reducing its variance. For further details on how to divide observations in training and validation-test sets in econometrics and ML see Varian (2014). The main idea is about averaging a set of (de-correlated) trees resulting from *B* different bootstrapped samples reducing variance[2]. The trees are estimated considering a random subset *m* of the *p* predictors[3] in order to stop the models from being too similar and, consequently, correlated. When the problem is continuous, the prediction is represented by the average of the *B* estimated trees, whereas for qualitative variables the outcome is selected among *B* classes according to majority vote. See Biau and Scornet (2016) for further details on random forests.

2 The underlying idea is rather trivial. In fact, such an approach relies on the fact that the variance of the mean sample (for *iid* elements) is σ^2/B so that, as the number of estimated models increases, the variance goes to zero. Trouble arises if the average models are similar themselves, i.e. there are similar combinations of predictors and cut-points (*j*, *s*) in which the predictors' space is partitioned so that the obtained models are highly correlated.

3 For classification $m \approx \sqrt{p}$. See Hastie et al. (2009).

4 Dataset description

Our data spans from 2010 to 2018. Primary data come from the business register of the Chamber of commerce of Bolzano (Italy) containing a wide range of information at the firm level. In detail, we have historical demographic data about more than 200.000 firms that were recorded at the business register of Bolzano since the early 20th century and with at least a local unit in the same province. Among those, more than 58,000 were active between 2012 and 2018. For each firm and precise local unit localization, demographic events including registration, activation, closure and deletion from the register are recorded by date. Moreover, unstructured textual data describing further events concerning the firm are available. For more than 150,000 firms, it is known the NACE code describing which economic activity is (or was) performed and in which local unit, summing up to more than 300,000 local units including branches and offices. Additional textual descriptions of the performed activity are directly provided by the firm to the register and further textual information for craft businesses are available. As fare as individuals or legal entities involved in the firm are concerned, the dataset contains information about the shareholders, including their share of ownerships, their demographic details and their province of residence[11]. Information about members of boards and committees, corporate officers, their role in the firm and their appointment period are available too.

Our reference statistical units are firms activated between 2012 to 2016; start-ups' performances are tracked from the creation up to three years ahead. The reference period is just apparently short, if compared with data availability: in fact, detailed lagged information on firms' ownership, necessary for the matching process distinguishing real start-ups from regenerated firms, are available since 2010. In fact, three outcomes are identified for the purposes of the analysis as our response variables: 1) the survival of the firm for the first three years of life, 2) the cessation of the firm within three years and 3) the regeneration of the firm during its life cycle. The latter consists in the apparent death of the firms within 3 years after the activation and the consequent re-activation. Those events are identified based on a matching algorithm suggested by Eurostat (2007). Accordingly, a business birth is the creation of a combination of production factors with the restriction that no other enterprises are involved in the event while the death is the dissolution of a combination of factors of production. Firms' localization, economic activity and legal unit are compared to detect continuity linkages signaling the involvement of another enterprise in a demographic event (only) apparently occurring at the beginning or at the end of a firms' life cycle; rather, those companies - formally receiving a new ID number from the business register, or declaring to cease their activity- are actually connected to other entities, pre-existing or forthcoming.

We initially considered 41 regressors to predict firms' outcome. Regressors are collected and described in Table 1 and Table 2, distinguishing among environmental, idiosyncratic and industrial determinants of firms' survival as in Box (2008).

Table 1: Regressors, their class and definition.

Risk Source	Variable Label	Class	Description
Economy specific input	Unemployment T	Numeric	Unemployment rate at the month T in which the firms is activated.
Economy specific input	Unemployment T-1	Numeric	Unemployment rate at the month T-1.
Economy specific input	Unemployment T-6	Numeric	Unemployment rate at the month T-6.
Economy specific input	Unemployment T-12	Numeric	Unemployment rate one year before the activation.
Economy specific input	Unemployment d T-1	Numeric	Variation in the unemployment rate: 1 month's lag.
Economy specific input	Unemployment d T-6	Numeric	Variation in the unemployment rate: 6 months' lag.
Economy specific input	Unemployment d T-12	Numeric	Variation in the unemployment rate: 12 month's lag.
Economy specific input	Unemployment T+3	Numeric	Lag of three months, forward.
Economy specific input	ST Interest rate T	Numeric	Short therm interest rate at the month T in which the firm is activated.
Economy specific input	ST Interest rate T-1	Numeric	Short therm interest rate at the month T-1.
Economy specific input	ST Interest rate T-6	Numeric	Short therm interest rate at the month T-6.
Economy specific input	ST Interest rate T-12	Numeric	Short therm interest rate one year before the activation.
Economy specific input	ST Interest rate d T-1	Numeric	Variation in the interest rate: 1 month's lag.
Economy specific input	ST Interest rate d T-6	Numeric	Variation in the interest rate: 6 months' lag.
Economy specific input	ST Interest rate d T-12	Numeric	Variation in the interest rate: 12 month's lag.
Economy specific input	ST Interest rate T+3	Numeric	Lag of three months, forward.
Economy specific input	Spread T	Numeric	Spread between Italian 10-y BTP and German 10-y bonds yields at the month T in which the firm is activated.
Economy specific input	Spread T-1	Numeric	Lagged spread between Italian 10-y BTP and German 10-y bonds.
Economy specific input	Spread T-6	Numeric	Lagged spread between Italian 10-y BTP and German 10-y bonds.
Economy specific input	Spread T-12	Numeric	Lagged spread between Italian 10-y BTP and German 10-y bonds.
Economy specific input	Spread d T-1	Numeric	Lagged monthly variation of the spread.
Economy specific input	Spread d T-6	Numeric	Lagged variation of the spread.
Economy specific input	Spread d T-12	Numeric	Lagged yearly variation of the spread.
Economy specific input	Spread T+3	Numeric	Lag of three months, forward.
Firm specific input	Maturity	Numeric	Age of the pre existing firm (if any) at the time in which the startup is activated.
Firm specific input	Age	Numeric	Age of the individual controlliing the firm at the month in which it is activated.
Firm specific input	Average age	Numeric	Average age of idividuals involved in the firm.
Firm specific input	Family workers	Numeric	Number of workers from the family.
Firm specific input	Employees	Numeric	Number of non-familiar employees.
Firm specific input	Concentration	Numeric	Share of ownership of the main shareholder.
Firm specific input	Branches	Numeric	Number of local units belonging to the firm
Firm specific input	Legal form	Categorical	Legal form of the firm.
Firm specific input	Gender	Categorical	Gender of cotrolling individual (majority shareholder or highset administrative role). When the main shareholder is a legal entity, the variable assumes value "legal person".
Firm specific input	Citizenship	Categorical	Citizenship of the individual person controlliing the firm.
Firm specific input	Month of birth	Categorical	Month in which the firm is activated. It captures the effect of seasonality
Firm specific input	Municipality	Categorical	Town in which the firm has its headquarter. To reduce dimensionality, values are aggregated in "Bolzano", "Merano" and "Other".
Firm specific input	Type of control	Categorical	Distinguishing between individual or legal person controlling the firm.
Firm specific input	Cause	Categorical	Cause of death of the previously existing firm (if any).
Firm specific input	Regenerated	Dummy	Takes value equal to one if the apparently new firm is linked to a pre-existing one.
Firm specific input	P/A	Dummy	Indicates the presence of a potential principal/agent problem. The dummy takes value 1 when the main shareholder has residence in a municipality different from the one of the firm.
Industry specific input	Sector	Categorical	Aggregated NACE code at 2 digits level.

Regressors listed in Table 1 and 2 of the current section and Table 6 of Section 7 try to map the hypothesis stated across Section 2. A random forest is estimated in Section 6 using demographic and macroeconomic regressors as input variables, while accounting data available for a subsample are included to grow a conditional tree in Section 7. The predictive power of each dimension is assessed with "Mean-Decrease Gini"-based measures obtained from the random forest while the sign of the association between each input and the investigated output (the firms' performances) results from the split points of the conditional tree.

The role of family members (H1a) and their influence on firms' performances is proxied by the ratio between "Family workers" and the total number of employees (Table 1, firm-specific, numeric input). The categorical variable related to the "Legal form" and the numeric variable referred to the "Concentration" of the ownership are adopted to test H1b. Hypotheses H2a-H2d concern, respectively, the firms' size, their attitude toward investments on R&D, their capitalization and their legal form; in order to assess the relative importance of those features, a mix of demographic regressors and accounting ratios are exploited in Section 6 and in Section 7: more specifically, the relevant demographic inputs are the number of "Branches" and the "Legal form" (Table 1) while the main accounting ratios are "Research and developments" (R.D) and "Own founds" (Table 6). The categorical variables labelled "Municipality" and "Sector" account for the group of hypotheses H3a-H3d, referred to the characteristics of the geographic area and the sector in which operates the firm. The fourth set of hypotheses is concerned with the typical measures of production factor costs, namely interest rate and real wage, as well as the unemployment rate resulting from the aggregate economic conditions (Table 1, "Economy specific" inputs). To such extent, the "Spread" and its lagged values were also included as real-time measures of the perceived country-specific risk[44]. Finally, the last hypothesis refers to the role played by the entrepreneurial experience on firm's chances to survive. An element of novelty in the present work is represented by the inclusion in the space of regressors of a dummy variable taking value 0 if the new activation is a real birth according to the definition provided by Eurostat (2007), and 1 if the firm is linked to a previously existing productive unit. The regressor was built up based on the matching algorithm implemented by Caterini (2018); the latter consists of pairwise comparisons among apparently different firms based on their NACE code, legal unit and localization. Moreover, the cause of death of the ancestor enterprise is included as a categorical variable. Looking for exact matches, data standardization and pre-processing are needed, so that missing NACE codes for 25.645 local units were partially imputed using textual descriptions of the economic activity provided by the entrepreneur at the business register to train a random forest capable of classifying, once fed with texts, the economic sector. The classification accuracy was estimated to be 70% in Caterini (2018); a similar strategy was proposed by Roelands et al. (2017) with web-scraped descriptions.

In Section 6 the predictive power of several regressors feeding the random forest is assessed. Real time economic variables such as unemployment, interest rate and spread are expressed monthly and lagged backward and forward according to the month of birth of the firm. Firm-level information refer to the start-up itself or to the previously existing firm -if any- which was rebirth in the form of

44 If compared with other economic variables such as the GDP, the spread presents the advantage to be expressed with a higher frequency well suited with our data on monthly businesses creations and deaths.

the new activation. A labelled training set was generated distinguishing between start-ups and regenerated firms, that is, those new activations linked to previously existing productive units.

As we know the exact date in which a firm is activated, we included monthly economic and financial variables such as the interest rate observed in the province of Bolzano, the regional unemployment rate and the spread index between Italian and German 10-years treasury bonds as real time proxies of the stance of the economy. [45] Moreover, we considered the backward lagged values of those variables in the month immediately preceding the birth, 6 and 12 months earlier, the variation of the variable between the period of birth and the lagged periods and, in order to allow some nowcasting, the forward value 3 months after the birth. We included demographic features that are firm's specific. The age of the controlling individual, proxying her experience, is computed ranking shareholders based on the amount of owned stocks: the controlling one is defined as the majority shareholder[4]. When shares are equally distributed among multiple shareholders, the one with the highest hierarchical position in the firm is considered. In cases in which there were not enough information on the composition of the proprietorship, or the shareholders were legal persons, we considered the age of the highest appointed officer. The average age of individuals leading the firm was considered, including owners and individuals appointed at key controlling positions. The concentration of ownership was defined as the amount of stocks held by the majority shareholder as in Gul et al. (2010). Employees, family workers and the number of branches proxied the dimension of the firm and the managerial style. We also considered the legal form, the type of control, and the gender of the controlling individual[5]. Citizenships of controlling individuals were aggregated by geographic area except for Italy and China, due to the huge weight of those two countries in the productive system. Moreover, European countries were distinguished according to their participation in the EU. Countries not existing longer are classified as "OTHER". Categorical regressors for which statistical units with missing data were few were assigned the label "MISSING"; instead, missing data related to the sector, the age, the average age and the number of workers were imputed. For regenerated firms, the cause of death of the previously existing productive entity, as identified by the matching process, was considered. If a firm was linked to several pre-existing ones, the cause of death was imputed based on a majority vote rule. Instead, if the previously existing firm was still active,

45 Lagged values of real-time regressors can lead to issues related to multicollinearity. This would be dramatic dealing with algorithms such as linear discriminant analysis, in which multicollinearity would prevent the estimation of the model itself. Dealing with random forest, a possible approach to address such a point would be applying standard criteria like BIC and AIC or estimating a preliminary random forest in order to remove regressors with negligible variable importance, as proposed in Chapter 3.

4 The controlling individual of a sole proprietorship firm is the entrepreneur.

5 We distinguished between males, females, and legal persons whenever we didn't have information on natural persons leading the firm and the majority shareholder was a legal person.

the label “ACTIVE” was assigned. Finally, real start-ups were assigned value “NONE” to the cause of death of the linked company (since it doesn’t exist).

Table 2: Classes of categorical regressors.

Categorical Variable Label	Classes
Gender	Male, Female, Legal person.
Citizenship Aggr.	Mediterrean Africa, Central Africa, Central EU, East EU, East Europe, Italy, Latin America, North America, Middle East, PRC, Rest of Asia.
Month of birth	From January to December.
Code Municipality	Bolzano, Merano and Other.
Type of control	Individual and Legal person.
Cause	Active, Adm (Administrative reasosn), CE (transfer), CH (Closure of the local unit), CZ (Cessation of any activity), Death, Def, (Default), LO (Firm's rent), Merge/Split, None, Others, SC (Dissolution), TN (Change of legal form).
Sector	AGR (Agricolture), ART (Art and entertainement), BKG (Banking and Insurance), DISTR (Distribution), EDU (Education), ENG (Energy), EXTR (Extractions), HLTH (Health), ITC (Information and technology), MANUF (Manufacturing), OTHR (Others), PROC (Procurement), REST (Real estate), SC (Science and techniques), TRD (Trade), TRM (Tourism).

Table 3.a: Composition of the sample by year and sector

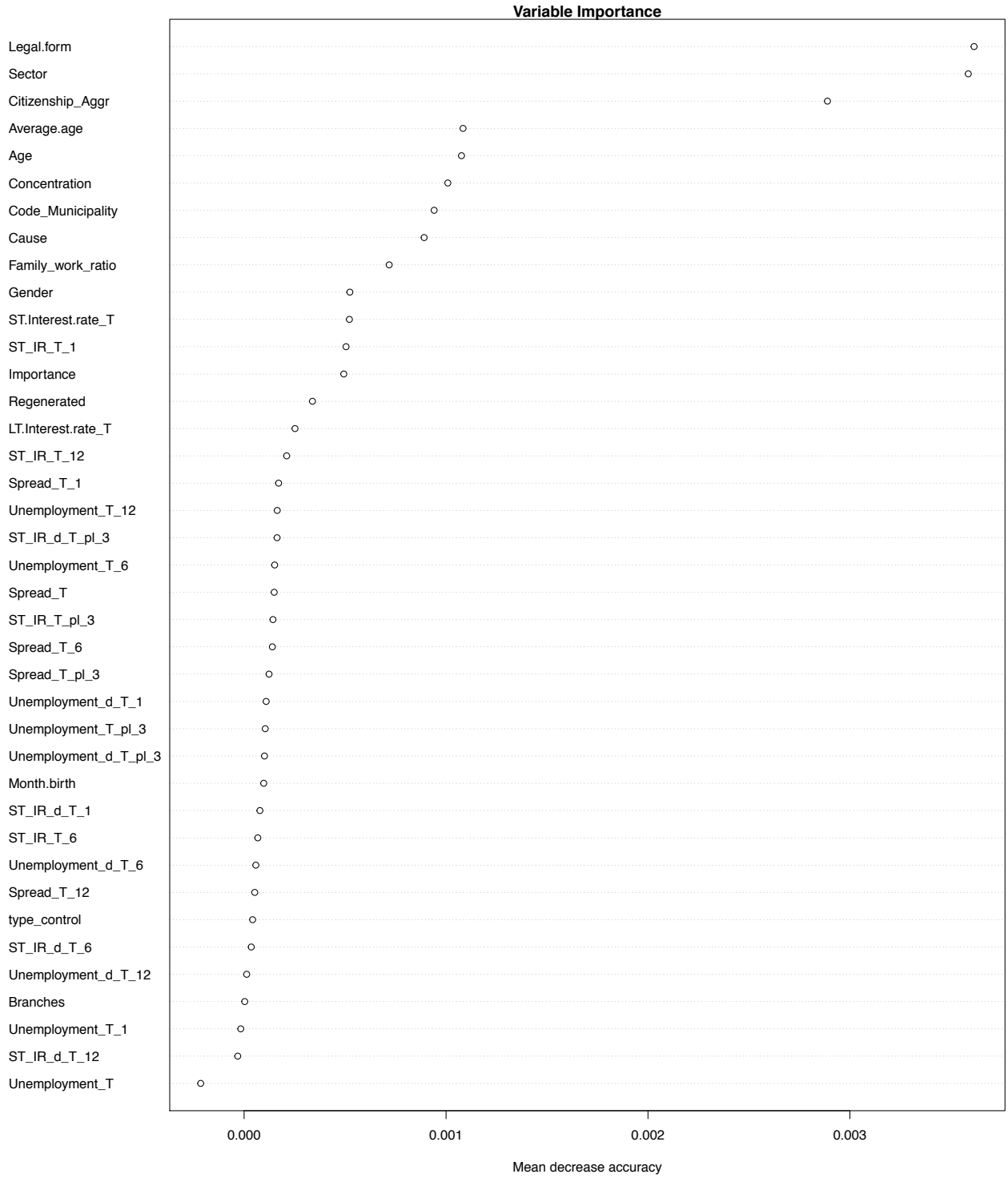
Aggregate Sector	2012	2013	2014	2015	2016	Total
AGR	439	424	356	320	380	**1919**
ART	18	31	29	20	32	**130**
BKG	25	56	25	36	34	**176**
DISTR	67	70	42	46	65	**290**
EDU	14	17	18	17	14	**80**
ENG	91	33	30	34	32	**220**
EXTR	0	0	0	1	0	**1**
HLTH	12	7	11	13	17	**60**
ITC	76	85	66	104	91	**422**
MISSING	5	0	0	4	0	**9**
MANUF	132	124	142	160	124	**682**
OTHR	93	101	98	103	93	**488**
PROC	1	0	0	2	0	**3**
REST	397	427	355	422	411	**2012**
SC	196	195	179	189	210	**969**
TRD	471	484	489	417	478	**2339**
TRM	461	424	400	454	485	**2224**
Total	**2498**	**2478**	**2240**	**2342**	**2466**	**12024**

Table 3.b: Composition of the sample by year and response

Response	2012	2013	2014	2015	2016	Total
Dead	345	339	326	291	269	**1570**
Regenerated	114	93	66	84	88	**445**
Surv_3	2039	2046	1848	1967	2109	**10009**
Total	**2498**	**2478**	**2240**	**2342**	**2466**	**12024**

6 Survival, death and regeneration with random forest and demographic regressors

A random forest with 300 (conditional) trees was estimated. Regressors' Gini-based scores are shown below. Important regressors are those for which the gain in terms of accuracy is high when including them in the model or, alternatively, when the exclusion from the set of randomly chosen regressors leads to a mean decrease in accuracy.



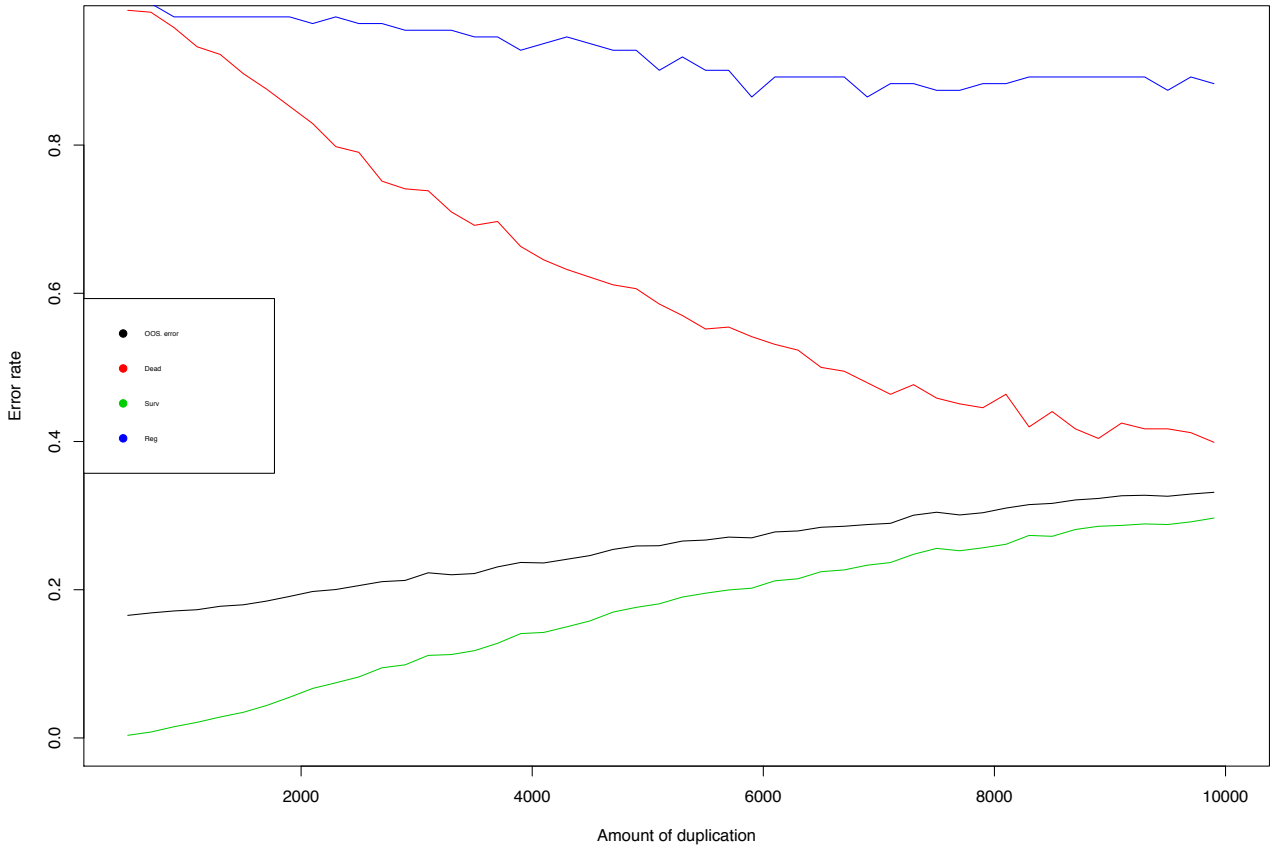
As it is shown in Table 4, a classification error of 16.5% is just partially satisfying: in fact, since the firms ceasing before the third year are the 16.7% of the sample, the classification performance achieved by the estimated algorithm is a negligible improvement respect to the assignment of each firm to the class "surv_3". This is the consequence of having an unbalanced sample in which the large majority of observations belong to a single class. Nevertheless, the researcher can decide which type of error she is more willing to tolerate[6]. In order to balance the trade-off between accuracy and sensitivity in a more plausible way, we perform some duplication on the classes that are less frequent in the training set, that is "Regenerated_LC" and "Dead", as suggested by Härdle et al. (2009); nonetheless, rather than setting the three classes to be equally represented in the training set, we suggest to compare different strategies of balancing until a satisfactory result is reached.

Table 4: Classification Performances

	Dead	Regenerated_LC	Surv_3	Total	Error rate
Dead	33	0	1539	1572	**97,90%**
Regenerated_LC	0	0	448	448	**100,00%**
Surv_3	1	0	10022	10023	**0,01%**
Total	**34**	**0**	**12009**	**12043**	**16,51%**

The figure below reports the classification errors for each class as the number of duplications increase. We divided the initial sample in a training set (75% of observations) and a test set (the remaining 25%); from a subset of observations belonging to the classes which needed to be balanced, we extracted (with replacement) random instances adding them to the training set and repeating the operation obtaining different compositions of the data. As we are more interested in predicting deaths, we duplicated the examples of the class "Dead" three times more than the observations belonging to the class "Regenerated_LC".

6 A bank which has to decide for the concession of a loan more likely prefers to misclassify a potential survivor rather than a potential loser.

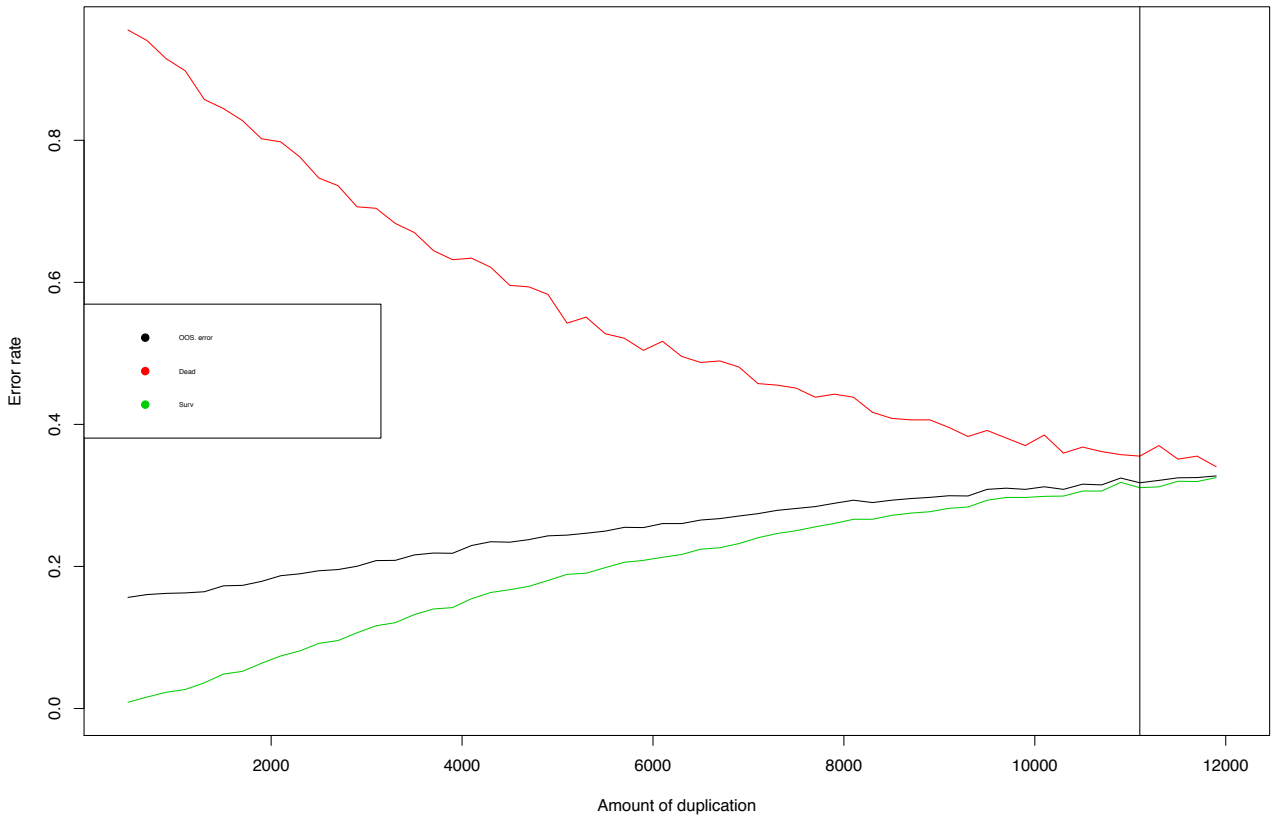


As it is shown in the graph above, duplicating the class for deaths (x-axis) we obtain a reasonably good balance between the overall misclassification rate, the misclassification rate for startups survived at least three years and for deaths (y-axis). Nevertheless, classification performances are still unsatisfactory for the regenerated firms, with a rather stable trend for their error rate (the blue line). As is shown in Table 5, when the firms regenerated during the life cycle are about the 17% of the training set, deaths are the 50%, survivals are about the 34%, and the test set maintains the original proportions among classes, regenerated firms are distributed barely reflecting the weights of the classes in the training set. The underlaying economic intuition is the following: independently on the composition of the data, the model systematically fails in recognizing regenerated firms as those firms are, to some extent, hybrids between deaths and survivors; in fact, companies that are regenerated during their life cycle are formally death but, practically speaking, they will restart their activity under a new shape. Consequently, in order to improve classification performances, we estimate a new random forest defining just two classes.

Table 5: Classification Performances after balancing

Outcome	Dead	Regenerated_LC	Surv_3	Total	Erorr rate
Dead	232	17	137	386	**39,90%**
Regenerated_LC	56	13	42	111	**88,29%**
Surv_3	693	53	1768	2514	**29,67%**
Total	**981**	**83**	**1947**	**3011**	**33,15%**

The new definition of the output variables is the following: firms which are regenerated within the first two years of life are labelled as "Dead" rather than regenerated, while firms that apparently die during their third year of life and are reactivated (according to the above matching algorithm) are included in the cluster of "Surv_3". Below is reported the variation in OOS error rate, the misclassification rate for deaths and survived firms as the amount of balancing increase. The vertical line indicates a fair enough balance between OOS error rate (31.7%), the error rate for deaths (35.5%) and for survived firms (31%). Those error rates are consistent with the ones obtained by Chen et al. (2011) applying vector machine (SVM) to accounting ratios to predict survival probabilities of German firms. As we already discussed, random forests have the advantage to show a ranking of variables' predictive power but are poor when economic interpretation is taken into account, as well as SVM. In the next section we will argue that estimating classification trees we gain a lot in terms of interpretability with an acceptable loss in terms of accuracy.



7 Survival and death including accounting ratios

Random forests are extremely useful for predictions. Moreover, variables' importance based on Gini index can be exploited to extract insights on the inputs' predictive power. Nevertheless, those models, as well as other machine learning tools such as SVM or artificial neural networks, are poor in assessing the direction of causal relations, if any[7]. Although trees are less reliable than forests due to their tendency to overfit the data, nonetheless they graphically represent which regressors have more predictive power and at which split-points, making them more suitable for descriptive purposes. Furthermore, tree-based methods outperform standard econometric linear models whenever there are nonlinearities and interactions (Varian, 2014). In this section we add to the space of predictors accounting data available between 2012 and 2016 for a subset of 2.473 firms over 12.024 to estimate a conditional tree; moreover, some of the main regressors according to the ranking obtained in Section 6 were retained, in detail: a group of regressors proxying the experience of the entrepreneur and other individuals involved in the firm (namely "Age", "Average age") including the categorical variable for the cause of death of the previously existing firm ("Cause"), the legal form and the family work ratio describing relevant features of the analyzed entities, the categorical variable for the sector, the spread index and the short term interest rate at the time of birth. Each group of regressors reflects the different factors affecting firms' outcomes, that is the environment, the macroeconomic situation and firms-specific factors. Financial regressors were partially defined according to accounting ratios exploited by Chen et al. (2011), subject to data availability. Tables 6 and 7 provide respectively the full set of accounting inputs and the relative descriptive statistics.

7 In fact, when economists estimate linear or logistic regressions, they are implicitly assuming a causal relation that exists just by construction, since our models are to the real data generating process as our world is to Plato's Ideas.

Table 6: Descriptive statistics for the main ratios

Variable description	Regressor name
EBITDA-depreciation and amortization	**EBIT**
Earning before taxes depreciation and amortization	**EBITDA**
Net profits	**Net.Profit**
Financial assets	**Financial.Assets**
Total financial assets	**Total.Fin.Assets**
Intangible+Tangible+Financial Assets	**Int.Tan.Fin.Assets**
Total Assets	**Total.Assets**
Plants Costs	**Plants.Costs**
Research and developements	**R.D.**
Patents	**Patents**
Licences	**Licences**
Goodwill	**Goodwill**
Cash	**Cash**
Circulating Assets	**Circulating.Assets**
Risk founds	**Appropriated.Risk.Funds**
Total debts	**Total.Debts**
Own founds	**Own.founds**
EBITDA over sales, profitability	**EBITDA/SALES**
Return on sales, profitability	**ROS**
Return on investments, profitability	**ROI**
Return on equity, profitability	**ROE**
Return on assets, profitability	**ROA**
Net profits/Revenues, profitability	**Profit_Margin**
EBIT over total assets, profitability	**EBIT/TA**
EBITDA over total assets, profitability	**EBITDA/TA**
EBIT over sales, profitability	**EBIT/SALES**
Bank loans over revenues	**Bank_Loans/Revenues**
Debts over total assets, leverage	**DEBTS/TA**
Cash over total assets, liquidity	**CASH/TA**
Total assets over sales, asset turnover	**TA/SALES**
log(Revenues)	**DIMENSION**
Wages over P (based on CPI in Alto Adige)	**W/P**

Table 7: Descriptive statistics for the main ratios.

	n	mean	sd	median	min	max	range	skew	kurtosis
EBITDA/SALES	1,759	- 7	101	6	- 884	498	1,382	- 4	27
ROS	1,263	1	15	2	- 49	30	79	- 1	2
ROI	812	2	12	1	- 30	30	60	- 0	0
ROE	1,665	16	53	8	- 149	141	290	- 0	0
ROA	2,110	- 5	55	1	- 979	180	1,159	- 6	69
Profit_Margin	2,115	- 64	2,908	-	- 133,603	5,071	138,674	- 46	2,099
EBIT/TA	2,114	- 1	20	0	- 929	2	931	- 46	2,098
EBITDA/TA	2,114	- 0	20	0	- 929	2	931	- 46	2,099
EBIT/SALES	2,114	- 66	2,904	0	- 133,496	872	134,368	- 46	2,106
Bank_Loans/Revenues	1,050	6	15	-	-	99	99	3	10
DEBTS/TA	2,114	1	29	1	-	1,329	1,329	46	2,105
CASH/TA	2,114	0	0	0	- 1	2	3	2	4
TA/SALES	2,115	3,959	169,217	2	0	7,781,983	7,781,983	46	2,108
DIMENSION	2,115	4	3	4	- 7	14	21	- 0	- 1
W/P	2,111	1	15	-	-	665	665	44	1,988

As this section is concerned with assessing the predictive power of a regressor and, eventually, describing its asymmetric effect according to the regions obtained by recursive binary splitting, we don't distinguish among training and test set; as a matter of fact, this exercise is aimed at providing policy implications devoted to professionals and policymakers in order to support new births after identifying their potential strengths and weaknesses.

Conditional trees are grown by recursive binary splitting: at each stage, the null hypothesis of independence between the output variable and the set of inputs is performed and the regressor mostly associated with the response is chosen; then, binary splits are performed until classification performances are improved, as exposed in Section 2. Conditional trees are *pre-pruned*, mining that pairs of regressors and split-points are chosen iteratively until the null hypothesis of independence among inputs and outputs is not rejected. Moreover, conditional trees handle missing observations performing surrogate splits as in Hothorn et al. (2006), allowing us to use the full dataset, although accounting data are available scarcely.

The resulting tree should be interpreted as follows: edges are ordered hierarchically according to variables' (decreasing) predictive power. Branches indicate split-points defining sub-regions occurring when nonlinear relations are detected. The values at which the splits occur should not be interpreted "literally", as numeric regressors were standardized to prevent different scaling affecting the final result. Finally, the grey rectangles indicate the number of observations belonging to each sub-region and the estimated probabilities to cease and to survive (first and second numeric values, respectively).

Consistently with the ranking of variables obtained with the random forest and with H1b, the legal form has a strong predictive power for firms' survival in the conditional tree. Not surprisingly, the algorithm distinguishes between sole proprietorships and other legal forms. As a matter of fact, the latter are more likely to survive as their survival probability is almost always greater than 90% (terminal nodes 18, 20 and 21). For them, the ROA (the second layer on the right-hand side of the tree) shows a high predictive power. Not surprisingly, the higher the ratio, the higher the probability to survive (always more than 92% in the right-hand side branches). Instead, when ROA is below a given threshold, firms' survival probability is affected by the presence of family-workers: the lower the ratio, the higher the proportion of survivals (ellipse number 14, terminal node number 15). For those firms with higher ROA, the dimension matters, defined as the logarithm of revenues at the first year of birth. Still, the greater the dimension, the higher the probability to survive (rectangles 20 and 21 have predicted response always greater than rectangle 18). The fact that grater firms are (subtly) more likely to survive is an extremely powerful result: in fact, since we are dealing with start-ups, no reverse causality is faced with reference to firm's dimension. In other words, analysing mature firms we face the risk to overestimate the importance of firms' dimension for their survival, as firms

necessarily have to survive in the previous periods in order to grow, so that growth is possible just if survival occurs and not necessarily viceversa. Instead, as our statistical units are start-ups, the dimension is the starting point rather than the arrival. Moreover, the cause of death of the pre-existing firm (if any) and the dummy variable indicating that the new activation is not a real birth, were not found to be predictive enough for the side of the tree that we are dealing with. Finally, as dimensions increase, consortiums, cooperatives, partnerships and joint stocks companies with a lower ratio between family workers and total employees are marginally more likely to survive, which contradicts H1a (ellipse number 19). Nevertheless, marginal differences in our results should not be neglected, as the 84% of the sample is composed by survivals, with the consequence that the death of a firm is a rather unlikely event and a 1% higher risk of cessation (0.055 against 0.068) should be accounted for carefully. Nodes 1 and 17, once considered jointly, provide evidence in favour of H1b and H2d, and against H2a.

As far as sole proprietorships are concerned, the sector in which the firm operates is predictive of its outcome. The agricultural sector ("Ag" in the tree) is clearly separated from the others and agricultural firms have, overall, greater chances to survive, that seems to be coherent with H3d. In agriculture, the age of the individual leading the firm is predictive of firm's survival probability: the younger, the better. For younger individuals, the cause of death of the previously existing firm, if any, has informative power too. Indeed, firms that were regenerated for administrative reasons or because the previous owners died (the case, for instance, of succession) are less likely to survive, while firms laying along the edge number 6 are likely more experienced since they emerged from demographic events such as split-offs, dissolutions, transfers, or were linked to firms still active. This result is made weaker by the presence, in leaf number 6, of start-ups that were not generated from previously existing firms (CAUSE is "none"): nonetheless, in our sample, the percentage of survivals among real start-ups (almost 82%) is lower than the percentage of survivals among regenerated firms (almost 85%), providing some evidences (although week) on the role of experience on firms' performances. The argument exposed above advocates for the importance of distinguishing between real start-ups and apparent births when addressing research questions related to firms' performances, success factors and liability of newness, as stated by H5. To finalize, node number 10 provides evidence in favour of H4b: the lower the interest rate, the higher the chance to survive.

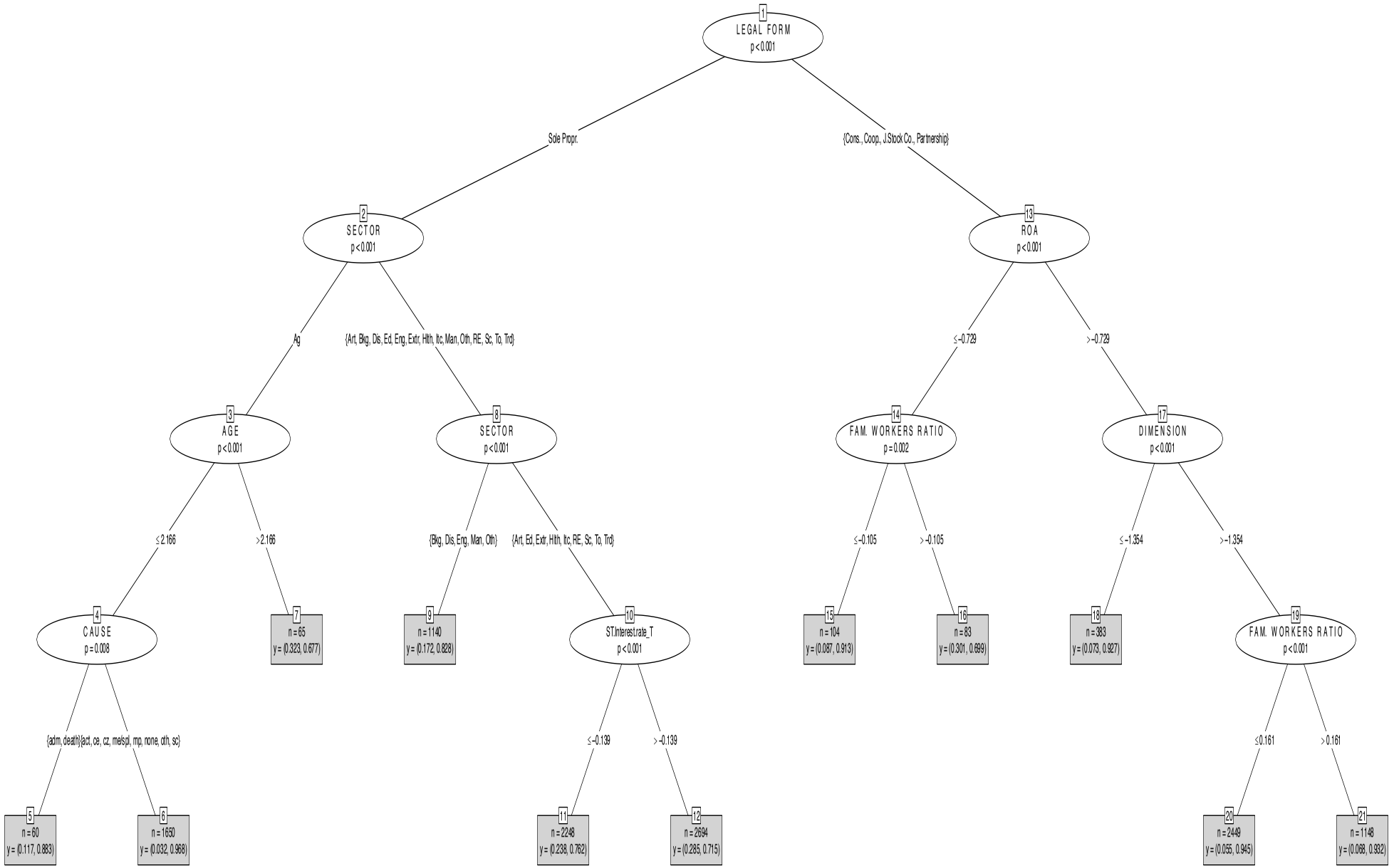
6 Conclusions

Classification problems are typical tasks in machine learning, whom algorithms reveal to be extremely powerful, especially when multiple outputs are dealt with and when the dataset is large, that is, several regressors are considered at the same time. This work represents, at the best of our knowledge, the first attempt to test jointly a wide range of hypothesis, exposed in Section 2, and to

apply a classification algorithm distinguishing between start-ups and regenerated firms. A random forest was estimated to predict start-ups outcomes, defined as dead, regeneration during the life cycle and survival to the third year. A sample of more than 12.000 firms activated in the province of Bolzano between 2012 and 2016 was used to feed the learning algorithms aimed at predicting survival chances, providing a ranking of variables' predictive power and describing, through the classification tree, the estimated sign of relevant variables. With the random forest we have been able to predict firm's outcome with an accuracy around 30%, showing that balancing the dataset and duplicating the under-represented observations the classification performances improve. The mean decrease in Gini index was used to provide a ranking of variables' predictive power. Accounting ratios were then combined to firm's specific, industrial and macroeconomic regressors in order to obtain descriptive insights on features' predictive power and to visualize their relevance regions. From variables importance we can confirm that H1b and H2d are verified, that is, the legal form matters, and proprietorship's concentration has negative effect on survival chances. Moreover, survival rates are negatively correlated with the presence of family workers in firms not owned by a sole proprietor. The sector results to be predictive and the survival rate in agriculture is higher with respect to other sectors when the sole proprietorship is considered. The municipality in which the firm operates matters too: we distinguished between urban area and rural ones and we don't have evidences to confirm H3a. Unemployment doesn't appear to be predictive: the reason can lie in the fact that unemployment rates were always extremely low in the considered region between 2012 and 2016. Instead, the short-term interest rate is negatively correlated with firms' survival according to the conditional tree. From the tree itself emerges the importance of ROA and dimensions, both positively correlated with firms' survival, while there are no evidences on the role of real wages and R&D investments, consistently with H2b. Finally, we have proved that, while dealing with start-ups, the information about previous entrepreneurial experiences matters, since more experienced firms have more chances to survive and the dummy variable identifying apparent births is ranked higher, for instance, than real time information on the stance of the economy.

Those results present the weakness of poor generalizability because of the nature of the sample; moreover, accounting data are available just for a subset of firms. Nevertheless, this work highlights the importance of distinguishing among real births and regenerated firms when analyzing start-ups success factors. Random forests result to be reliable in predicting firms' outcome, although few instances of dead and regenerated firms in the sample mine the prediction accuracy and the learning ability. As a matter of fact, random forests and conditional trees proved to be valid tools in identifying start-ups fragilities, providing to policy makers and professionals useful hints to support companies at the beginning of their life cycle.

CLASSIFICATION TREE FOR SURVIVALS



References

Abatecola, G., Uli, V. (2016), Entrepreneurial competences, liability of newness and infant survival, *Journal of Management Development*, Vol. 35 Iss 9, pp. 1082 – 1097.

Abatecola, G. (2014). Research in organizational evolution. What comes next?. *European Management Journal*, Vol. 32, No. 3, pp. 434-443.

Abatecola G., 2013. Survival or Failure within the Organizational Life Cycle. What Lessons for Managers?. *Journal of General Management*, 38(4): 23-38.

Abatecola G. (2012). Interpreting Corporate Crises: Towards a Co-Evolutionary Approach. *Futures*, 44: 860-869.

Abatecola, G., Cafferata, R., and Poggesi, S. (2009), "Revisiting Stinchcombe's "liability of newness": a systematic literature review", *International Journal of Globalisation and Small Business*, Vol. 3 No. 4, pp. 374-92.

Agarwal, R., Sarkar, M. B., & Echambadi, R. 2002. The conditioning effect of time on firm survival: An in- dustry life cycle approach. *Academy of Management Journal*, 45: 971–994.

Aldrich, H.E. (1999). Organizations evolving. London: Sage Publications.

Aldrich, H.E. and Auster, E.R. (1986), "Even dwarfs started small: liabilities of age and size and their strategic implications", *Research in Organizational Behavior*, Vol. 8 No. 2, pp. 165-98.

Amankwah-Amoah, J., Boso, N., Antwi-Agyei, I. (2016). The Effects of Business Failure Experience on Successive Entrepreneurial Engagements: An Evolutionary Phase Model, *Group & Organization Management*, Vol. 43, No. 4, 648-682.

Arribas, I., & Vila, J. (2007). Human capital determinants of the survival of entrepreneurial service firms in Spain. *International Entrepreneurship and Management Journal*, 3(3), 309-322.

Aspelund, A., Berg-Utby, T. and Skjevdal, R. (2005), "Initial resources' influence on new venture survival: a longitudinal study of new technology-based firms", *Technovation*, Vol. 25 No. 11, pp. 1337-1347.

Audretsch, D. (1995). Innovation, growth and survival. *International Journal of Industrial Organization*, 13(4), 441-457.

Audretsch, D. B., E. Santarelli, and M. Vivarelli, 1999. Start Up Size and Industrial Dynamics: Some Evidence from Italian Manufacturing, *International Journal of Industrial Organization* 17, 965–983.

Bakker, R. M., Josefy, M. (2018). More than just a Number? The Conceptualization and Measurement of Firm Age in an Era of Temporary Organizations. *Academy of Management Annals*, Vol. 12, No. 2,

Barnett, W.P. and Amburgey, T.L. (1990), "Do larger organizations generate stronger competition?", in Singh, J.V. (Ed.), Organizational Evolution: New Directions, Sage, Newbury Park, CA, pp. 78-102.

Bates, T. (2005). Analysis of young, small firms that have closed: delineating successful from unsuccessful closures. *Journal of Business Venturing*, 20(3), 343-358.

Bercovitz, J., & Mitchell, W. 2007. When is more better? The impact of business scale and scope on long-term business survival, while controlling for profitability. *Strategic Management Journal*, 28: 61–79.

Bermiss, Y. S., & Murmann, J. P. 2015. Who matters more? The impact of functional background and top executive mobility on firm survival. *Strategic Management Journal*, 36: 1697–1716.

Boeker, W. 1988. Organizational origins: Entrepreneurial and environmental imprinting of the time of founding. In G. R. Carroll (Ed.), Ecological models of organizations: 33–51. Cambridge, MA: Ballinger.

Box, M. (2008). The death of firms: Exploring the effects of environment and birth cohort on firm survival in Sweden. *Small Business Economics*, 31(4), 379-393.

Breslin, D. (2016). What Evolves in Organizational Co-Evolution? *Journal of Management and Governance*, 20(1) 45-67.

Breslin, D. (2011). Interpreting futures through the multi-level co-evolution of organizational practices. *Futures*, Vol. 43, pp. 1020-1028.

Bruderl, J., & Schussler, R. 1990. Organizational mortality: The liabilities of newness and adolescence. *Administrative Science Quarterly*, 35: 530–547.

Byrne, O., & Shepherd, D. A. (2015). Different strokes for different folks: Entrepreneurial narratives of emotion, cognition, and making sense of business failure. Entrepreneurship Theory and Practice, 39, 375-405.

Cader, H. A., & Leatherman, J. C. (2011). Small business survival and sample selection bias. *Small Business Economics*, 37(2), 155-165.

Cafferata, R. 2016. Darwinist connections between the systemness of social organizations and their evolution, *Journal of Management & Governance*, Vol. 20 No. 1, pp. 19-44.

Cafferata, R., Abatecola, G., Poggesi, S. (2012), "Arthur Stinchcombe's "liability of newness": contribution and impact of the construct", *Journal of Management History*, Vol. 18 No. 4 pp. 402. 418.

Campbell, N. D., Heriot, K. C., Jauregui, A., & Mitchell, D. T. (2012). Which state policies lead to US firms exits? Analysis with the Economic Freedom Index. *Journal of Small Business Management*, 50(1), 87–104.

Caterini, G. (2018). Classifying Firms with Text Mining (No. 2018/09). Department of Economics and Management.

Caves, R. E. (1998). Industrial Organization and New Findings on the Turnover and. *Journal of economic literature*, 36(4), 1947-1982.

Chen, S., Härdle, W. K., & Moro, R. A. (2011). Modeling default risk with support vector machines. Quantitative Finance, 11(1), 135-154.

Cope, J. (2011), "Entrepreneurial learning from failure – an interpretative phenomenological analysis", *Journal of Business Venturing*, Vol. 26 No. 6, pp. 604-623.

Cristofaro, M. (2017). Countervailing the Liability of Newness by Bringing in Active Initial Investors: The Case of Facebook. *Strategic Direction*, Vol. 33, No.8, pp. 1-3.

Darwin, C.R. (1859), On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life, John Murray, London.

Deakin, E.B. (1972), "A discriminant analysis of predictors of business failure", *Journal of Accounting Research*, Vol. 10 No. 1, pp. 67-179.

Dias, A., Teixeira, A.A.C. (2017). The anatomy of business failure: A qualitative account of its implications for future business success. *European Journal of Management and Business Economics*, Vol. 26 No. 1, pp. 2-20.

Dibrell, C., Craig, J.B., Moores, K., Johnson, A.J., Davis, P.S. (2009) Factors Critical in Overcoming the Liability of Newness: Highlighting the Role of Family, *The Journal of Private Equity*, Vol. 12, No. 2, pp. 38-48.

Delgado, M., Porter, M., & Stern, S. (2010). Clusters and entrepreneurship. *Journal of Economic Geography*, 10(4), 495-518.

Dyke, L., Fisher, E., & Reuber, A. (1992). An inter-industry examination of the impact of owner experience on firm performance. *Journal of Small Business Management*, 73-86.

Ejermo, O., & Xiao, J. (2014). Entrepreneurship and survival over the business cycle: how do new technology-based firms differ? *Small Business Economics*, 43(2), 411-426.

Eggers, J. P., & Song, L. 2015. Dealing with failure: Serial entrepreneurs and the costs of changing industries between ventures. *Academy of Management Journal*, 58: 1785–1803.

Esteve-Pérez, S., & Llopis, J. (2004). The determinants of survival of Spanish manufacturing firms. *Review of Industrial Organization*, 25(3), 251-273.

Esteve-Perez, S. and Manez-Castillejo, J.A. (2008), "The resource-based theory of the firm and firm survival", *Small Business Economics*, Vol. 30 No. 3, pp. 231-49.

Everett, J. and Watson, J. (1998), "Small business failure and external risk factors", *Small Business Economics*, Vol. 11 No. 4, pp. 371-390.

Fackler, D., Schnabel, C., Wagner, J. (2013). Establishment exits in Germany: the role of size and age. *Small Business Economics*, Vol. 41, pp. 683-700.

Fotopoulos, G. and H. Louri, 2000, Determinants of Hazard Confronting New Entry: Does Financial Structure Matter? *Review of Industrial Organization* 17(3), 285–300.

Fichman, M. and Levinthal, D.A. (1991), "Honeymoons and the liability of adolescence: a new perspective on duration dependence in social and organizational relationships", *Academy of Management Review*, Vol. 16 No. 2, pp. 442-68.

Folta, T., Cooper, A., & Baik, Y. (2006). Geographic cluster size and firm performance. *Journal of Business Venturing*, 21(2), 217-242.

Freeman, J., Carroll, G.R. and Hannan, M.T. (1983), "The liability of newness: age dependence in organizational death rates", *American Sociological Review*, Vol. 48 No. 5, pp. 692-710.

Geroski, P. A., Mata, J., & Portugal, P. 2010. Founding conditions and the survival of new firms. *Strategic Management Journal*, 31: 510–529.

Gimeno, J., Folta, T., Cooper, A., & Woo, C. (1997). Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. *Administrative Science Quarterly*, 42(4), 750- 783.

Giovannetti, G., Ricchiuti, G., & Velucchi, M. (2011). Size, innovation and internationalization: a survival analysis of Italian firms. *Applied Economics*, 43(12), 1511-1520.

Hodgson, G.M. and Knudsen, T. (2004), "The firm as an interactor: firms as vehicles for habits and routines", *Journal of Evolutionary Economics*, Vol. 14 No. 3, pp. 281-307.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Hsu, D. K., Wiklund, J., & Cotton, R. D. 2015. Success, failure, and entrepreneurial reentry: An experimental assessment of the veracity of self-efficacy and prospect theory. *Entrepreneurship Theory and Practice*, 41: 19–47.

Jang, J., and Danes, S.M. (2013). Are We on the Same Page?: Copreneurial Couple Goal Congruence and New Venture Viability, *Entrepreneurship Research Journal*, Vol. 3, No. 4, 483-504.

Jenkins, A. S., Wiklund, J., & Brundin, E. 2014. Individual responses to firm failure: Appraisals, grief, and the influence of prior failure experience. *Journal of Business Venturing*, 29: 17–33.

Josefy, M.A., Harrison, J.S., Sirmon, D.G., & Carnes C. 2017. Living and dying: synthesizing the literature on firm survival and failure across stages of development. *Academy of Management Annals*, 11(2): 770–799.

Justo, R., DeTienne, D. R., & Sieger, P. 2015. Failure or voluntary exit? Reassessing the female under- performance hypothesis. *Journal of Business Venturing*, 30: 775–792.

Kalnins, A., & Williams, M. 2014. When do female-owned businesses out-survive male-owned businesses? A disaggregated approach by industry and geography. *Journal of Business Venturing*, 29: 822–835.

Mantere, A., Aula, P., Schildt, H. and Vaara, E. (2013), "Narrative attributions of entrepreneurial failure", *Journal of Business Venturing*, Vol. 28 No. 4, pp. 459-473.

Mata, J. and P. Portugal, 1994, Life Duration of New Firms, *Journal of Industrial Economics* 42, 227–246.

Mata, J., & Portugal, P. (2002). The survival of new domestic and foreign owned firms. *Strategic Management Journal*, 23, 323-343.

Mata, J. and P. Portugal, 1999, Technology Intensity, Demand Conditions, and the Longevity of Firms, in D.B. Audretsch and A.R. Thurik (eds.), Innovation, Industry Evolution and Employment, Cambridge, UK: Cambridge University Press, pp. 265–279.

Mata, J., Portugal, P., & Guimaraes, P. (1995). The survival of new plants: Start-up conditions and post-entry evolution. *International Journal of Industrial Organization*, 35,607-627.

Michael, S.C. and Sung, M.K. (2005), "The organizational ecology of retailing: a historical perspective", *Journal of Retailing*, Vol. 81 No. 2, pp. 113-23.

Nelson, R.R. and Winter, S.G. (1982), An Evolutionary Theory of Economic Change, Harvard University Press, Cambridge, MA.

Platt, H. & Platt, M. (1994). Business cycle effects on state corporate failure rates. *Journal of Economics and Business*, 46, 13-127.

Porter, M. (1990). The comparative advantage of nations. The Free Press, New York .

Renski, H. (2015). Externalities or Experience? Localization Economies and Start-up Business Survival. *Growth and Change*, Vol. 46 No. 3, pp. 458-480.

Revilla, A.J Pérez-Luño, A., Nieto, M.J. (2016). Does Family Involvement in Management Reduce the Risk of Business Failure? The Moderating Role of Entrepreneurial Orientation, *Family Business Review*, Vol. 29, No. 4, pp. 365-379.

Salman, A.K., von Friedrichs, I., Shukur, G. (2011). The determinants of failure of small manufacturing firms: assessing the macroeconomic factors, *International Business Research*, Vol. 4, No. 3, 22-32.

Schutjens, V., & Wever, E. (2000). Determinants of firm success. *Papers in Regional Science*, 79, 135-153.

Segarra, A., & Callejón, M. (2002). New firms´survival and market turbulence: New evidence from Spain. *Review of Industrial Organization*, 20(1), 1-14.

Shepherd, D.A., Williams, T.A. and Patzelt, H. (2015), "Thinking about entrepreneurial decision making: review and research agenda", *Journal of Management*, Vol. 41 No. 1, pp. 11-46.

Sorenson, O., & Audia, P. (2000). The social structure of entrepreneurial activity: Geographic concentration of footwear production in the United States. *The American Journal of Sociology*, 106(2), 424-462.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American economic review*, 71(3), 393-410.

Stinchcombe, A.L. (1965), "Social structure and organizations", in March, J. (Ed.), Handbook of Organizations, Rand McNally, Chicago, IL, pp. 142-93.

Strottmann, H. (2007). Entrepreneurial Survival, *Small Business Economics*, Vol. 28, pp. 87-104.

Sutton, 1987. The process of organizational death: Disbanding and reconnecting. *Administrative Science Quarterly*, 32: 542–569.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28.

Varuma, V. C., Rocha, C.A. (2012). The effect of crises on firm exit and the moderating effect of firm size. *Economic Letters*, Vol. 114, pp. 94-97.

Walsh, G.S. and Cunningham, J.A. (2016), "Business failure and entrepreneurship: emergence, evolution and future research", *Foundations and Trends in Entrepreneurship*, Vol. 12 No. 3, pp. 163-285.

Wennberg, K., & Lindqvist, G. (2010). The effect of clusters on the survival and performance of new firms. *Small Business Economics*, 34(3), 221-241.

Wholey, D.R., Christianson, J.B. and Sanchez, S.M. (1992), "Organizational size and failure among health maintenance organizations", *Administrative Sociological Review*, Vol. 57, pp. 829-42.

Williams, D.E. (2014). Resources and Business Failure in SMEs: Does Size Matter?, *Journal of Business and Management*, Vol. 20, No. 2, 89-102.

Wilson, N., Wright, M., & Scholes, L. 2013. Family business survival and the role of boards. *Entrepreneurship Theory and Practice*, 37: 1369–1389.

Yamakawa, Y., Peng, M.W. and Deeds, D.L. (2015), "Rising from the ashes: cognitive determinants of venture growth after entrepreneurial failure", *Entrepreneurship Theory and Practice*, Vol. 39 No. 2, pp. 209-236.

Zheng, W. T., Singh, K., & Mitchell, W. (2015). Buffering and enabling: The impact of interlocking political ties on firm survival and sales growth. *Strategic Management Journal*, 36: 1615–1636.