Analysis of Web Data for the Purpose of Predicting SMEs Financial Growth

Completed Research Paper

Introduction

Small and medium enterprises (SMEs) play an important role in the economy of many countries. When the overall world economy is considered, SMEs represent 95% of all businesses in the world, accounting for 66% of the total employment and 55% of the total production (OECD 2004). However, existing studies show that the current business environment is characterized as highly turbulent, influenced by modern information and communication technologies, globalization and employee mobility (Antlová 2009; Post 1997). Additionally, the growing number of SMEs caused competition to become increasingly intensive, forcing SMEs to experience more severe challenges in maintaining their existence and expanding their business.

Given the importance of SMEs to economy and society (Nooteboom 1998), public policy makers and academic researchers have put efforts in helping to trigger growth in SMEs, ultimately enhancing overall economic performance (Carter and Van Auken 2006). Thererfore, predicting the growth of SMEs has become an important area of research (Davidsson and Klofsten 2003; Pompe and Bilderbeek, 2005).

In order to support SMEs at improving their competitiveness, researchers and academics have been analyzing factors influencing the success of SMEs for many decades (Altman 1968; Ohlson 1980; Henebry 1996). Moreover, with the emergence of big data, researchers turned their focus on applying data mining techniques to build risk and growth prediction models (Kim and Sohn 2010; Duman et al. 2012; Kruppa et al. 2013). However, current prediction models only include few data types such as financial or operational data and thus cannot explain the whole and complex context of SMEs growth (Ozgulbas and Koyuncugil 2012). Moreover, conventional data collection is primarily conducted via questionnaire studies, which is very laborious and time-consuming, or provided by financial institutes, thus not publicly available and highly sensitive to privacy issues. Furthermore, data mining techniques such as artificial neural network and decision tree are extensively studied with a strong focus on risk evaluation and bankruptcy prediction for SMEs. Although numerous studies on SMEs growth factors and growth modelling exist, studies reporting data mining based SMEs growth prediction cannot be identified.

Recently, web mining (WM) has emerged as a new approach towards obtaining valuable business insights. WM enables an automated and large-scale collection and analysis of potentially valuable data from the web such as the national commercial register and company websites. While WM methods have been frequently studied to anticipate growth of sales volume for e-commerce platforms (Patel et al. 2011), their application for assessment of SMEs growth factors is still scarce. To date, only few researchers studied the application of WM methods for SMEs growth prediction (Antlová et al. 2011; Li et al. 2016). Considering the large and increasing amount of data freely available online, WM bears a great potential in revealing valuable information hidden in web, which can be further used to build SMEs growth prediction models.

In this work, we explore the potential of using web data for SMEs growth prediction. In the first step, a large set of well-studied growth factors is identified from an extensive literature review, which will serve as a groundwork to generate the input for the SMEs growth modelling. Next, WM methods are applied to automatically collect and extract growth factors from various web data sources. Finally, we build a binary classification model using a supervised machine learning algorithm. More specifically, the developed model classifies a firm either in a non-growing or growing firm. In particular, the Random Forest Classifier (RFC) is used (Breiman 2001), which share a predominant role in a range of research domains (Cutler et al. 2007).

The remainder of this paper is structured as follows. First, we provide an overview of the previous work in related research areas. This contains an overview of SME research including SMEs growth model and growth factors, followed by a survey of data mining and WM studies in the domain of SME research. Based on the literature review, we identify the research gap addressed by this work and formulate the research

questions. Next, we provide an overview into the applied methodology. Consequently, we present the results and discuss our findings with regard to the previously identified research gap. This paper concludes with a summary and an outlook on future research.

Related Work

This section reviews the background of SMEs growth research by starting with the definition of SMEs and growth, followed by a brief overview of the factors influencing the growth of SMEs. Next, a summary of SMEs growth prediction modelling is provided. Finally, previous research on WM techniques for SMEs growth prediction is summarized and the identified research gap are pointed out.

Definition of SMEs and Growth

The definition of SMEs varies quite widely from country to country and even within single countries, depending on the business sector concerned. The World Business Council for Sustainable Development report (2007) stressed that there is no universally agreed definition of SMEs. According to the commission of the European union, SMEs are non-subsidiary, independent firms which employ fewer 250 employees. Additionally, the turnover of SMEs should not exceed EUR 50 million, and balance sheet total should be less than 43 million € (Potter and Storey 2007).

There is no consistency in the dimension of growth which theorists have used as the object of analysis. Different definitions have been used in the studies that attempted to explain the growth of SMEs (Delmar et al. 2003; Barringer et al. 2005; Delmar and Wiklund 2008). Non-financial growth measures include growth of employment, customer satisfaction and loyalty, whereas financial growth measures include growth of revenues, profits and assets. In this study, the adopted definition of growth is revenue growth, due to its importance to the economy (Lev et al. 2010).

Survey of Factors influencing SMEs Growth

The current business environment is characterized as complex and fast-changing, influenced by a variety of firm internal and external factors as illustrated in Figure 1 (Worthington and Britton 2009). Firm-internal factors can be roughly divided into two groups: (1) the characteristics of the firm such as firm attributes (age, size, location) and firm strategies (marketing, training strategies), and (2) the characteristics of the entrepreneur such as socio-demographic characteristics (age, gender, family and educational background) and the personality of the entrepreneur (need for achievement, risk-taking propensity). Firm-external factors can be divided into 2 groups: factors reflecting (1) the immediate and (2) the contextual environment. The immediate environment includes supplier and customer relationship, competition, labor market and resource market. In contrast, the contextual environment comprises macro-environmental factors such as economic, political, socio-cultural, technological and legal influences on the growth of businesses which can emanate not only from local and national sources but also from international developments (Worthington and Britton, 2009). In this study, we identified over 40 factors influencing the growth of SMEs through an extensive literature review, covering the whole and complex business environment (Appendix).

Survey of SMEs Growth Prediction Studies

The literature on business growth models dates back to 1967 and has proliferated since then into different streams addressing specific industries and business sizes. Lippitt and Schmidt (1967) developed a general growth model for all sizes of businesses by examining how personality development theories influences the creation, growth and maturation of a business. A few years later, Steinmetz (1969) qualitatively analyzed the growth of small enterprises by partitioning the growth curve of small enterprises into different stages and assessing the characteristic attributes of each stage. A qualitative study conducted by Scott and Bruce (1987) suggested a model for small business growth supporting managers to plan for future growth. The proposed model isolates five growth stages characterized by an unique combination of firm attributes. As a small company goes through different growth stages, attributes such as the management style, organizational structure and the use of technology changes. Although stage models are widely accepted among researchers and practitioners, stage models are criticized on some counts (O'Farrell and Hitchens

1988). First, some of them seem more than heuristic classification schemes rather than a conceptualization of the processes underlying growth. Second, they implicitly assume that a small business will either grow and pass through all stages or fail in the attempt. Empirical evidence does not justify such an assumption. Third, the models only include firm internal characteristics such as management style and organizational structure and do not incorporate environmental influences on the business growth. Furthermore, empirical research are conducted on small sample sizes and specific types of businesses via questionnaires studies and thus, could threaten the validity of stage models (Farouk and Saleh 2011).

With the emergence of big data, data mining techniques have been extensively studied in the domain of SME research. However, these studies mostly focused on the prediction of SME risk evaluation and bankruptcy rather than SMEs growth modelling. An early study indicated that backpropagation neural network were the most popular machine learning techniques among researchers in the finance and business domain during the 1990's (Wong et al. 2000). For instance, Zhang et al. (1999) provide a comprehensive review of Artificial Neural Network (ANN) applications for bankruptcy prediction. Their findings indicated that ANN perform significantly better than logistic regression models. West (2000) investigated the credit scoring accuracy of various ANN models in comparison with traditional methods such as logistic regression and discriminant analysis model. Consistent with the findings of Zhang et al. (1999), ANN models perform slightly better than traditional methods.

Other techniques widely applied in the domain of risk evaluation and bankruptcy prediction includes decision trees (DT) and their ensemble variations such as random forests (RF). For instance, Fantazzini and Figini (2009) developed a model based on RF for SME credit risk measurement and compared its performance with the traditional logistic regression approach. They came to the conclusion that both models provided similar results in terms of performance, highlighting the potential of RF for credit risk modelling. Another recent and more application-oriented study conducted by Ozgulbas and Koyuncugil (2012) proposed an early warning system based on DT-algorithms for SMEs to detect risk profiles. The proposed system uses financial data to identify risk indicators and early warning signs, and create risk profiles for the classification of SMEs into different risk levels.

In summary, data mining techniques such as ANN, DT and RF are extensively applied for SME risk evaluation and bankruptcy prediction. However, even though numerous studies on SMEs growth factors and models exist, research reporting data mining based SMEs growth prediction cannot be identified. Furthermore, current prediction models usually include one type of data sources and thus cannot explain the whole and complex context of SMEs growth (Ozgulbas and Koyuncugil 2012).

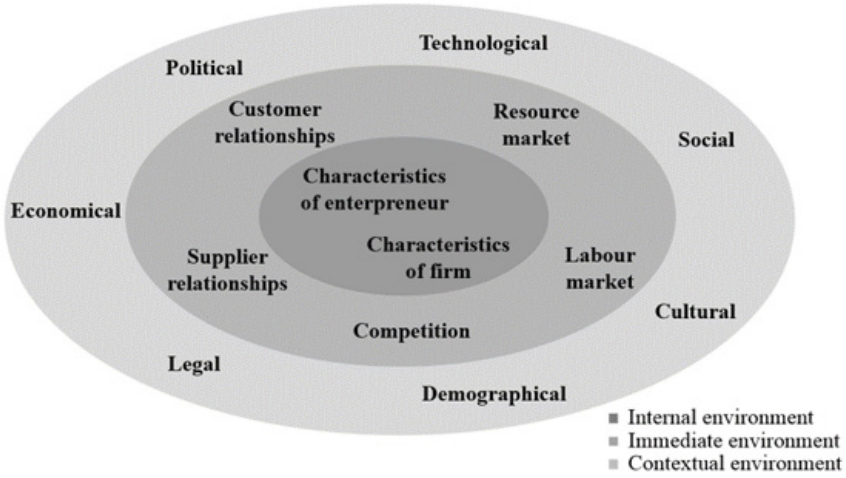

Figure 1. Factors influencing SMEs Growth.

Survey of Web Mining Based SMEs Growth Prediction Studies

The web is a popular and interactive medium with intense amount of data freely available for users to access. It is a collection of documents, text files, audios, videos and other multimedia data [28]. With billions of web pages available in the web, it is a rapidly growing key source of information, presenting an opportunity for businesses and researchers to derive useful knowledge out of it. However, automatically extracting targeted or potential valuable information from the web is a challenging task because of many factors such as size of the web, its unstructured and dynamic content as well as its multilingual nature. Therefore, WM research has emerged for knowledge discovery from the web.

Web Mining (WM) was coined by Etzioni (1996), to denote the use of data mining techniques to automatically discover Web documents, extract information from Web resources and uncover general patterns on the Web. WM research overlaps with other areas such as artificial intelligence along with machine learning techniques, data mining, informational retrieval, text mining and Web retrieval. WM research is classified on the basis of two aspects: the retrieval and the mining. The retrieval focuses on retrieving relevant information from large repository whereas mining research focuses on extracting new information already existing data (Sharda and Chawla). In general, WM tasks can be classified into three categories (Kosala and Blockeel 2000): Web Content Mining, Web structure mining and Web usage mining. WM has been proved very useful in the business world, especially in e-commerce. In the increasingly fierce competition in the e-commerce, any information related to consumer behavior are extremely valuable. A major challenge of e-commerce is to understand customers' needs and value orientation as much as possible, in order to ensure competitiveness in the e-commerce era. Thus, WM can be used to find data which have potential value from the website of e-commerce companies. For instance, Lin and Ho (2002) proposed a system, which extracts content information from a set of web pages with a goal of extracting the informative content. Morinaga et al. (2002) presented a system for finding the reputation of products from the internet to support marketing and customer relationship management in order to increase the sales growth. Finally, Thorleuchter et al. (2012) analyzed the impact of textual information from e-commerce companies' websites on their commercial success by extracting web content data from the most successful top 500 worldwide companies. The authors demonstrated how WM and text mining can be applied to extract e-commerce growth factors from the websites.

While WM methods has been well researched and used in the field of e-commerce research to increase the sales growth, it has barely been applied for SMEs growth research. The study conducted by Antlová, et al. (2011) is one of the first and few studies that demonstrated the potential of WM for SMEs growth prediction. In their paper, they examined the relationship between long-term growth of SMEs and a web presentation. They applied WM techniques to automatically extract potential valuable information for growth prediction. Recently, Li et al. (2016) explored micro-level characteristics and impacts of external relationships such as government or university relations on the growth of SMEs by extracting business-relevant information from company websites through WM, highlighting the potential of applying WM for SMEs growth research.

However, these studies only focus on the information available in company websites and thus, restrict the amount and spectrum of growth factors to the information typically given in company websites (Gök et al. 2015). Hence, further research exploiting the full potential of web data for SMEs growth prediction is required.

Research Gap and Research Questions

As pointed out in the previous subsections, several research gaps are identified in the domain of data mining and web mining based SMEs growth prediction modelling. First, most of the prediction models for SMEs focus on the anticipation of credit risk or bankruptcy. Although numerous studies on SMEs growth models exist, research reporting data mining based SMEs growth modelling are rare. Moreover, these models only include a few number of growth-influencing factors, which cannot capture the whole mechanism of SMEs growth. Second, conventional data collection to assess the growth factors is primarily conducted via questionnaire studies, which is very laborious and time-consuming. Furthermore, questionnaire studies suffer from well-known pitfalls such as low response rates and response bias (Arora et al. 2013; Lussier and Halabi 2010). Third, studies in the direction of using web data for SMEs growth prediction modelling are very limited. Additionally, most studies only focus on the information given in company websites (Arora et al. 2013; Gök et al. 2015; Li et al. 2016).

In order to address these issues, in this work we analyze the use of publicly available web data for SMEs growth prediction with the goal of addressing the following research questions:

RQ1. Which growth factors can be extracted from the web?

RQ2. From all the potential growth factors, which one prove to be discriminative?

RQ3. Can we solely use web data for SMEs growth prediction?

The following section details the methodology applied in order to obtain the answers.

Methodology

Data Collection

In this study, two different types of data are collected: (1) not publicly available data provided by a large Swiss insurer, and (2) publicly available web data. The data provided by the Swiss insurer serve as a ground-truth, i.e. labeled data for supervised machine learning. The web data are collected to derive the growth-indicating factors, which serve as input features to train a growth prediction model.

Ground truth Data

The data provided by the Swiss insurer contain information of a large set of Swiss SMEs, which consists of the company's name and the annual revenue in the period from 2014-2016. Thus, the data are used two-fold: (1) as a ground truth to train the growth model by constructing the growth label from the revenue data, and (2) as a linkage to collect company-related data from the web via company's name. In total, data of 1424 Swiss SMEs are collected for the purpose of this study.

Web Data

Web data related to the set of Swiss SMEs with known revenues (i.e. ground truth) are collected and factors influencing growth are extracted by means of web mining techniques. First, the usability of various web data sources is manually inspected with respect to the identified growth factors, as summarized in the appendix. In this study, five web data sources are examined.

- **Central Business Names Index (CBNI)**: CBNI provides free access to basic company information and links through to internet excerpts from the individual canton commercial registry databases. The freely viewable information for each company includes: company name, Swiss-wide identification number, registration date, legal form, address, purpose, status, and information about the members of the administrative board and their work function.
- **Federal Institute of Intellectual Property (FIIP)**: FIIP is the national registry for intellectual property of Switzerland (Federal Institute of Intellectual Property 2016). The FIIP database is publicly accessible and contains information about the registered patents and brands of companies. The counts of the registered patents and brands are collected and used as a measure for the innovativeness of a company, which is an important attribute of the firm characteristics (Ashton & Sung 2006).
- **Open Street Map (OSM):** OSM is a free-to-access web-based mapping system for navigation, location-based services and general information (OSM 2016). In this study, two types of datasets are downloaded from the OSM database of Switzerland in June 2016: (1) the Point of Interest (POI) dataset and (2) the Roads dataset. POIs are specific point locations on a map that might be deemed as useful or interesting for specific activities. They are described by the latitude and longitude or address of the location, type, name and contain 7 types of POIs: gastronomy (restaurant, bar, café), healthcare (hospital, pharmacy, doctor), public buildings (post, police, bank, school, university), public transportations (bus, tram, taxi and train station), tourism (museum, attraction, gallery), entertainment (cinema, theatre, casino, arts center, nightclub), parking lots and residential area. The Roads dataset contains 6 types of roads: motorway, trunk roads, primary road, secondary road, tertiary road and unclassified roads, which are described by the latitude and longitude of the nodes spanned across the roads. These datasets are used to derive factors reflecting the infrastructure surrounding the SMEs, which are proven to be influential on SMEs growth (Bottasso and Conti 2010). Therefore, the company address of the ground truth data collected from the CBNI are geocoded, and the POIs and roads within a radius between 100m and 500m are extracted for each SME.

- **Swiss Federal Statistical Office (SFSO):** SFSO is the national service provider and competence center for statistical observations in areas of national, social, economic and environmental importance (Swiss Federal Statistical Office 2016). The FSO is the main producer of statistics in the country and runs the Swiss Statistics data pool, providing information on all subject areas covered by official statistics. The dataset include socio-demographic, cultural, economic and political factors describing the Swiss population. Many of these factors were deemed in past studies as significantly influencing the SMEs growth. The census data was derived from annual portraits provided by the SFSO (Swiss Federal Statistical Office: STAT-TAB 2016): population density, population change, foreign nationals, age pyramid (young, adult, and old population ratios), area usage (settled and used for agriculture/forests/unused ratios), unemployment rate, residential density (persons per apartment room), and the number of businesses and residents employed in the different economy sectors (primary, secondary, and tertiary sector ratios). All data is aggregated on the level of municipalities - the lowest administrative unit on which Swiss census data is publicly available.
- **Swiss Federal Tax Administration (SFTA)**: SFTA is the Swiss administration for taxation, which manages the cantonal and municipal tax regulations (Swiss Federal Tax Administration 2016). The Swiss taxation system is very complex, divided into many tax categories (Feld and Kirchgässner 2001). In this study, we focus on the collection of the corporate taxation, which has proven to influence the SMEs growth (Bartlett and Bukvič 2001). Therefore, we extracted two factors reflecting the corporate taxation: (1) the profit tax, based on the net profit as accounted for in the corporate income statement, and (2) the capital tax, which is levied on the ownership equity of companies. The tax data are provided on a cantonal level.

The data from OSM, SFSO and SFTA are downloaded in CSV and PDF format. However, the information in CBNI and FIIP are only visible. Therefore, web scraping is applied to automatically capture information contained in CBNI and FIIP (Mitchell 2015). The web data sources along with the extracted growth factors are summarized in Table 1.

Table 1. Web Data Sources and Extracted Growth Factors

Business environment	Factor type	Growth factor	Web data source
Characteristics of firm	Firm attributes	Age	CBNI
		Size	
		Location	
	Firm resources	Human capital	
	Organization structure	Work specialization	
		Centralization of work	
		Legal form	
	Firm strategies	Innovativeness	FIIP
Contextual environment	Technological	Infrastructure	OSM
	Economical	Taxation	SFTA
		Business type diversity	SFSO
		Employment status	
	Social-cultural	Social class	
		Lifestyle	
		Cultural diversity	
	Political-legal	Regulatory environment	
	Demographical	Population size	
		Population density	
		Age distribution	

Data Preprocessing

Growth Label Creation

It is part of the data mining procedure to define the proper label based on the business objective for the data mining analysis. In this study, we build a binary classification model for SMEs growth, i.e. separating SMEs into non-growing and growing companies. In the first step, we use the annual revenue of 2015 and 2016 to calculate the relative change in revenue based on the following formula (Baptista and Leitão 2015):

$$relative\_growth = \frac{revenue_{2016} - revenue_{2015}}{revenue_{2015}} \times 100\%$$

Figure 2 shows the distribution of the ground truth data as a function of the relative revenue growth in percent. Out of 1424 SMEs, 132 SMEs (9.27%) showed a negative revenue growth, whereas 1069 SMEs (75.07%) showed no signs of growth (i.e. *relative_growth* = 0), and 223 SMEs (15.66%) experienced a growth between 2015 and 2016. Since the primary interest of our study is to predict the growth of SMEs, the samples with negative revenue growth are removed from the dataset, because the underlying mechanism of "shrinking" companies differs from the one of non-growing or growing companies, and thus are not included in this study (Arasti et al. 2014). To construct the binary labels, SMEs showing no signs of growth are assigned the value 0, whereas growing SMEs are assigned the value 1. Finally, the dataset consists of 1069 SMEs labelled with 0 as the majority class (82.74%) and 223 SMEs labelled with 1 as the minority class (17.26%).

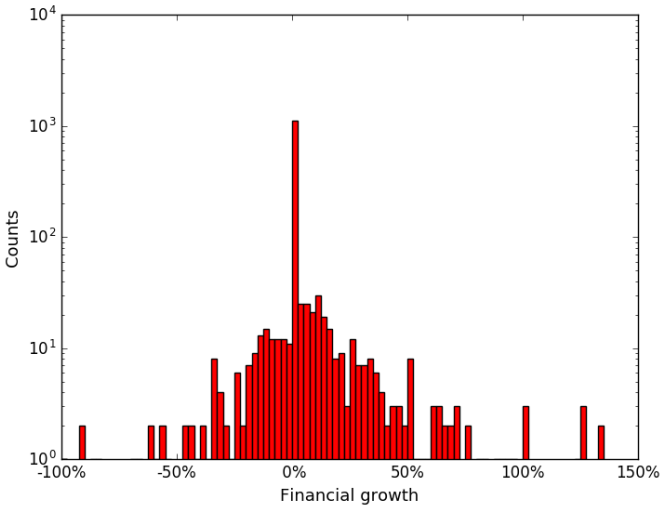

Figure 2. Distribution of the ground truth data (histogram bins = 100).

Input Feature Creation

The input features for growth modelling are derived from the collected web data as described above (Table 1). Whereas the information downloaded from OSM, SFSO and SFTA are provided in the form of structured numerical data and thus, require minimal data preprocessing, information extracted from the CBNI and FIIP by means of web mining are given in the form of semi-structured text data (i.e. html documents). Therefore, web content mining techniques are applied to extract the textual information reflecting the identified growth factors as shown in Table 1. Next, these information are preprocessed to numerical (e.g. firm size and age) and categorical values (e.g. firm legal form).

Because web data are often imperfect, the generated features are incomplete. To address this issue, missing attribute values are imputed with the most frequent attribute value (Grzymala-Busse and Hu 2001). Furthermore, features with zero variance and high correlation (*Pearson* correlation coefficient $r_{prs} \geq 0.97$) are removed (Hunt 1986). In total, 71 input features are generated for the purpose of supervised machine

learning, as summarized in Table 2. Note, that feature ID number 40 to 49 are dummy variables for the categorical feature "firm legal form".

Table 2. Input features for RFC

Feature ID	Feature name	Feature ID	Feature name
1	Population	37	Firm work specialization
2	Population density	38	Firm human capital
3	Foreigner	39	Firm work centralization
4	Population (0 to 19 years)	40	Firm legal form 1
5	Population (20 to 64 years)	41	Firm legal form 2
6	Population (over 64 years)	42	Firm legal form 3
7	Area (km^2)	43	Firm legal form 4
8	Area settlement	44	Firm legal form 5
9	Area agriculture	45	Firm legal form 6
10	Area woodland	46	Firm legal form 7
11	Area unproductive [%]	47	Firm legal form 8
12	Number of employed	48	Gastronomy businesses within 100m
13	Employed in primary sector	49	Healthcare businesses within 100m
14	Employed in secondary sector	50	Motorway within 100m
15	Employed in tertiary sector	51	Parking lots within 100m
16	Number of businesses	52	Pedestrian zones within 100
17	Businesses in primary sector	53	Public buildings within 100m
18	Businesses in secondary sector	54	Public transportation within 100m
19	Businesses in tertiary sector	55	Apartments within 100m
20	Housing ownership rate	56	Streets within 100m
21	Empty flats rate	57	Touristic attractions within 100m
22	Political votes for FDP	58	Entertainment businesses within 100m
23	Political votes for CVP	59	Gastronomy businesses within 500m
24	Political votes for SP	60	Healthcare businesses within 500m
25	Political votes for SVP	61	Motorway within 500m
26	Political votes for BDP	62	Parking lots within 500m
27	Political votes for EVP	63	Pedestrian zones within 500
28	Political votes for GLP	64	Public buildings within 500m
29	Political votes for PDA	65	Public transportation within 500m
30	Political votes for GPS	66	Apartments within 500m
31	Political votes for rightwing	67	Streets within 500m
32	Political votes for other groups	68	Touristic attractions within 500m
33	Profit tax	69	Entertainment businesses within 500m
34	Capital tax	70	Gastronomy businesses within 500m
35	Firm age	71	Healthcare businesses within 500m
36	Firm size		

Construction of the SMEs Growth Model

The growth of SMEs is a highly complex mechanism, thus predicting the growth of SMEs requires a machine learning algorithm which is capable to handle a high level of complexity. Therefore, we use the Random Forest Classifier (RFC), which is able to model complex interactions between the input variables and thus, share a predominant role in a range of research domains (Cutler et al. 2007). RFC is a non-parametric non-linear classification algorithm that uses an ensemble of decision trees, where each tree is built from a bootstrapped subset of the training data set (Breiman 2001). In constructing the ensemble, Random Forests use two types of randomness. First, each decision tree is based on a random subset of the observations and second, each split within each tree is created based on a random subset of features. This randomness results in low correlation of the individual trees, leading to the desirable properties of low bias and low variance (Hastie et al. 2005). Furthermore, a by-product of the random forest algorithm is the measure of feature importance, which allows a data-based evaluation of the relative importance of the factors for SMEs growth.

In order to build the RFC, a grid-search method was applied for finding optimal values for the hyper-parameters, which are (Boulesteix et al. 2012): (1) *n_estimators* - the number of trees grown in the RFC, (2) *max_features* - the number of features included to grow the individual trees, and (3) *min_samples_leaf* - the minimum sample leaf size, which is the end node of a decision tree. Furthermore, due to the imbalance of our ground truth data, a cost sensitive learning is introduced according to Breiman et al. (2004). Since the RFC tends to be biased towards the majority class, a heavier penalty on misclassifying the minority class is placed by assign a larger weight to the minority class (i.e. higher misclassification cost).

To report on classification performance of the RFC, we make use of accuracy scores. Accuracy is defined as the fraction of correctly classified samples of both, positive and negative classes. This makes it easy to interpret and ensures a neutral interpretation with respect to importance of positive and negative classes (Sokolova and Lapalme 2009). Additionally, in order to validate the classifier to the data, a stratified 10-fold cross-validation procedure was applied (Weiss and Kulikowski 1991). In a stratified 10-fold cross-validation, the original sample is partitioned into 10 subsamples while maintaining the ratio of the classes. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, while the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds is averaged to produce a single performance estimation (Kohavi 1995). In order to provide further insights on the classification performance, sensitivity and specificity scores are also reported together with the accuracy score (Sokolova and Lapalme 2009).

Results

To answer RQ1, a wide range of growth factors reflecting the internal and external business environment are identified and collected from governmental websites (CBNI, FIIP, SFSO, SFTA) and the OSM database (Table 1). In total, 71 input features are created to train the binary RFC for SMEs growth.

To answer RQ2 and RQ3, the following steps are performed. First, a 10-fold cross-validated grid search was conducted to find the optimal value for the parameters (*n_estimators, max_features, min_samples_leaf*), which were found to be (650, 0.79, 19). Next, based on the optimized RFC, we evaluate the explanatory power of the input features by reporting the feature importance, which is an inherent measure of the random forest algorithm (Breiman 2001). Figure 3 depicts the importance values for all 71 features. Large positive values indicate informative variables, whereas small values indicate predictors unlikely to be informative. From the plot, we see that "firm age" (feature ID 35) is clearly the most predictive feature with a substantially larger importance value than all other predictors, followed by a set of features reflecting the demographical, political, social-cultural and economical environment (features ranked from 2 to 14). The next features ranked from 15 to 31 are characterized by a mixture of features reflecting the firm characteristics and remaining contextual environment, with an emphasis on the features portraying the infrastructure surrounding the SMEs within a disk of 100m (feature ID 48 to 58). The legal form of firms (feature ID 40 to 47) and the infrastructure with a radius of 500m (feature ID 59 to 71) are not informative at all.

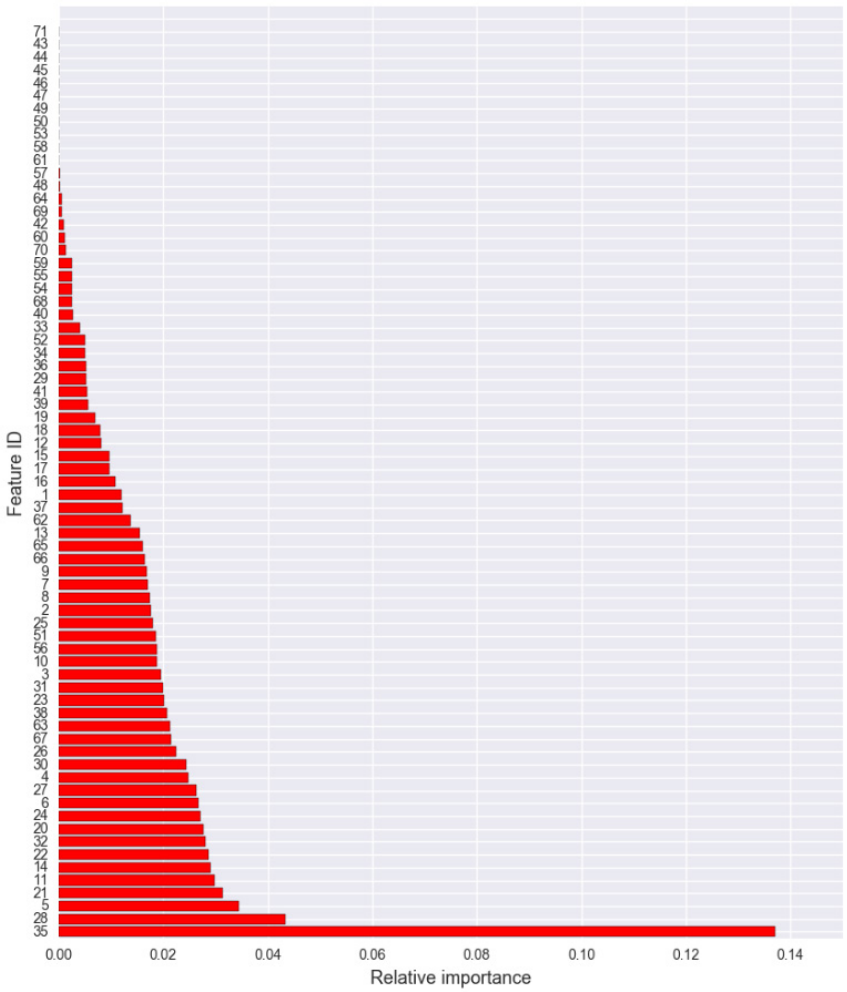

Figure 3. Feature Importance Plot.

Table 3 shows the classification performance of our RFC with respect to a binary classification of samples into non-growing and growing SMEs. The performance measures used are the accuracy - the overall percentage correctly classified, sensitivity - the fraction of samples correctly classified as growing SMEs, and specificity - the percentage of samples correctly classified as non-growing SMEs. The accuracy of the RFC is 60,2%. Our RFC favored specificity over sensitivity leading to a higher specificity score of 61.5% compared to the sensitivity of 53.0%.

Table 3. Classification Performance of Random Forest Classifier

Accuracy	60,2%
Sensitivity	53,0%
Specificity	61,5%

Discussion and Implications

In this paper we analyze the use of web data for the purpose of predicting the financial growth of SMEs. First, factors influencing the growth of SMEs are identified through an extensive literature review, as summarized in the Appendix. Next, a set of web data sources are examined with regards to the identified growth factors. Within the scope of this study, five web data sources containing information reflecting the business internal and external environment of SMEs are identified: Central Business Names Index (CBNI), Federal Institute of Intellectual Property (FIIP), OpenStreetMap (OSM), Swiss Federal Statistical Office (SFSO) and Swiss Federal Tax Administration (SFTA). The data are either downloaded from the websites (SFSO, SFTA, OSM) or collected by means of web scraping (CBNI, FIIP). Text mining methods are used to extract the growth factors from the scraped data (i.e. html documents) and to construct the input features for building the SMEs growth prediction model. Finally, a random forest classifier is trained on 71 features to predict a binary outcome, i.e. non-growing versus growing SMEs. The performance of our classifier is 60.2%, exceeding the performance of a random binary classifier (Neiberg and Laskowski 2006). This result highlights the potential of building a web data based growth prediction model for SMEs.

This study has both theoretical and practical implications. It contributes to the existing literature of SMEs growth research by confirming previous findings in a data-driven and model-based manner through machine learning. Furthermore, the proposed approach can be used to identify new growth factors and thus, extend the empirical body of knowledge.

Besides of the theoretical aspects, this study has a number of important practical implications. The built system allows an automated collection and analysis of publicly available web data in large scale with the objective of predicting future growth opportunities of SMEs. For the Swiss SME organizations, the insights generated in study may support the Swiss SME organizations at understanding the success and failure of SMEs, thus strengthen their supportive role for SMEs. For the investment companies, the built system can be used to monitor the development of SMEs by mining the changes in the internal and external business environment from web data, serving as an “early recognition system” for future opportunities of growth. Finally, for SMEs, the system can be used to evaluate the characteristics of firms based on the information given in the web. The absence of important key success factors can be pointed out to firms, thus serving as a consulting program.

Limitations and Future Work

This study is not without limitations and provides several opportunities for further research. First, our work is limited to Switzerland, thus the obtained results might differ in different geographical regions. In order to address this issue we plan to conduct our study with data from different countries in order to be able to make a generalization of the obtained results. Second, important growth factors completing the firm-internal environment, such as the business type and the characteristics of the entrepreneur (appendix) are not included in our model because they are not available in the examined web data sources. To address this issue, we plan to apply web mining techniques to collect and preprocess textual information given in company websites and social platforms like Xing, with the goal to enlarge the input feature space of our model. Finally, we plan to test different machine learning algorithms with to goal to optimize the performance of our model in predicting the SMEs growth.

Appendix

Business environment		Factor type	Growth factor
Internal environment	**Characteristics of firm**	Firm attributes	Age of firm (Nguyen et al. 2004) Size (Harabi 2003) Location (Liedholm 2002)
		Firm resources	Financial resources (Beck et al. 2006) Human capital (Wiklund et al. 2009)
		Firm strategies	Innovativeness (Ashton & Sung 2006) HR management (Huselid 1997)
		Organization structure	Work specialization (Olson et al. 2005) Centralization (Olson et al. 2005) Legal form (Olson et al. 2005)
	Characteristics of entrepreneur	Socio-demographical	Age of entrepreneur (Reynolds et al. 2000) Family background (Gray et al. 2006) Education (Brush et al. 2001) Experience (Robertson et al. 2003)
		Personality	Need for achievement (Barkham 1994) Locus of control (Mueller and Thomas 2001) Risk-taking propensity (Stewart et al. 2003)
		Competences	Managerial (Ibrahim and Goodwill 1986) Entrepreneurial (Wang and Ang 2004)
External environment	**Immediate environment**	Customer relationships	Customer needs (Reinartz et al 2004) Customer acquisition (Reinartz et al 2004) Customer retention (Reinartz et al 2004)
		Supplier relationships	Quality & cost (Pearson and Ellram 1995) Service reliability (Quale 2003)
		Competition	Clusters of competitors (Folta et al. 2006) Cluster of complementary firms (Delgado et al. 2014) Product pricing (Gadenne 1998)
		Labor market	Employment status (Michie and Sheehan 2005) Contract type (Michie and Sheehan 2005)
		Resource market	Natural resources (Judge and Douglas 1998) Energy resources (González-Benito 2005)
	Contextual environment	Technological	Information (Swierczek and Ha 2003) Infrastructure (Bottasso and Conti 2010)
		Socio-cultural	Social class (Gurol and Atsan 2006) Lifestyle (Walker and Brown 2004) Cultural diversity (Richard 2000)
		Economical	Financial resources (Beck et al. 2006) Taxation (Robertson et al. 2003)
		Political-legal	Government support (Yusuf 1995) Regulatory environment (Edwards et al. 2004)
		Demographical	Population size (Mazzarol et al. 1999) Population density (Mazzarol et al. 1999) Age & gender distribution (Mazzarol et al. 1999)

References

Altman, E. I. 1968. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The journal of finance* (23:4), pp. 589-609.

Antlová, K. 2009. "Motivation and Barriers of ICT Adoption in Small and Medium-Sized Enterprises," *E+M Ekonomie a Management* (22:2), pp. 140-154.

Antlová, K., L. Popelínsky and Tandler, J. 2011. "Long term growth of SME from the view of ICT competencies and web presentations," *E+ M Ekonomie a management* (4), p. 125.

Arasti, Z., Zandi, F. and Bahmani, N. 2014. "Business failure factors in Iranian SMEs: Do successful and unsuccessful entrepreneurs have different viewpoints?,".*Journal of Global Entrepreneurship Research* (4:1), pp. 1-14.

Arora, S. K., Youtie, J., Shapira, P., Gao, L., and Ma, T. 2013. "Entry strategies in an emerging technology: a pilot web-based study of graphene firms," *Scientometrics* (95:3), pp. 1189-1207.

Ashton, D. N. and Sung, J. 2002. *Supporting workplace learning for high performance working,* Geneva: International Labour Organization.

Baptista, R. and Leitão, J. 2015. *Entrepreneurship, human capital, and regional development*, Heidelberg: Springer. pp. 15-28.

Barkham, R. J. 1994. "Entrepreneurial characteristics and the size of the new firm: a model and an econometric test," *Small Business Economics* (6:2), pp. 117-125.

Barringer, B. R. and Harrison, J. S. 2000. "Walking a tightrope: Creating value through interorganizational relationships," *Journal of management* (26:3), pp. 367-403.

Bartlett, W. and Bukvič, V. 2001. "Barriers to SME growth in Slovenia," *MOCT-MOST: Economic Policy in Transitional Economies* (11:2), pp. 177-195.

Beck, T. and Demirguc-Kunt, A. 2006. "Small and Medium Size Enterprise: Access to finance as a growth constraint," *Journal of Banking and Finance* (30:1), pp. 2931-2943.

Beck, T., Demirguc-Kunt, A. and Maksimovic, V. 2006. "The influence of financial and legal institutions and firm size," *Journal of Banking and Finance* (30), pp. 2995-3015.

Bottasso, A. and Conti, M. 2010. "The productive effect of transport infrastructures: does road transport liberalization matter?," *Journal of Regulatory Economics* (38:1), pp. 27-48.

Boulesteix, A. L., Janitza, S., Kruppa, J. and König, I. R. 2012. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2:6), pp. 493-507.

Breiman, L. 2001. "Decision-tree forests," *Machine Learning* (45:1), pp. 5-32.

Breiman, L., Chen, C. and Liaw, A. 2004. „Using random forest to learn imbalanced data," *Journal of Machine Learning Research*, p. 666.

Brush, C., Greene, P. and Hart, M. 2001. "From initial idea to unique advantage: The entrepreneurial challenge of constructing a resource base," *Academy of Management Executive* (15:1), pp. 64-80.

Carter, R. and Auken, H. V. 2006. "Small firm bankruptcy," *Journal of Small Business Management* (44:4), pp. 493-512.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. 2007. "Random forests for classification in ecology," *Ecology* (88:11), pp. 2783-2792.

Davidsson, P. and Klofsten, M. 2003. "The business platform: Developing an instrument to gauge and to assist the development of young firms," *Journal of small business management* (41:1), pp. 1-26.

Delgado, M., Porter, M. E. and Stern, S. 2014. "Clusters, convergence, and economic performance," *Research Policy* (43:10), pp. 1785-1799.

Delmar, F., Davidsson, P. and Gartner, W. B. 2003. „Arriving at the high-growth firm," *Journal of business venturing* (18:2), pp. 189-216.

Delmar, F. and Wiklund, J. 2008. "The Effect of small business managers growth motivation on firm growth: a longitudinal study," *Entrepreneurship Theory and Practice* (32:3), pp. 437-457.

Duman, E., Y. Ekinci and Tanrıverdi, A. 2012. "Comparing alternative classifiers for database marketing: The case of imbalanced datasets," *Expert Systems with Applications* (39:1), pp. 48-53.

Edwards, P., Ram, M. and Black, J. 2004. "Why does employment legislation not damage small firms?," *Journal of Law and Society* (31:2), pp. 245-265.

Etzioni, O. 1996. "The World-Wide Web: quagmire or gold mine?," *Communications of the ACM* (39:11), pp. 65-68.

Fantazzini, D. and Figini, S. 2009. "Random survival forests models for SME credit risk measurement," *Methodology and Computing in Applied Probability* (11:1), pp. 29-45.
Farouk, A. and Saleh, M. 2011. "An Explanatory Framework for the Growth of Small and Medium Enterprises," In: *International Conference of System Dynamics Society*.
Federal Institute of Intellectual Property 2016. www.swissreg.ch/. *accessed 15 February 2017*
Feld, L. P. and Kirchgässner, G. 2001. „Income tax competition at the state and local level in Switzerland," *Regional Science and Urban Economics* (31:2), pp. 181-213.
Folta, T. B., Cooper, A. C. and Baik, Y. 2006. "Geographic cluster size and firm performance," *Journal of Business Venturing* (21:2), pp. 217–242.
Gadenne, D. 1998. "Critical Success Factors for Small Business: An Inter-industry Comparison," *International Small Business Journal* (17:1), pp. 36-56.
González-Benito, J. 2005. "Environmental proactivity and business performance: an empirical analysis," *Omega* (33:1), pp. 1-15.
Gök, A., Waterworth, A. and Shapira, P. 2015. "Use of web mining in studying innovation," *Scientometrics* (102:1), pp. 653-671.
Gray, K., Foster, H. and Howard, M. 2006. "Motivations of Moroccans to be entrepreneurs," *Journal of Developmental Entrepreneurship* (11:4), pp. 297-318.
Grzymala-Busse, J. and Hu, M. 2001. "A comparison of several approaches to missing attribute values in data mining," In: *Rough sets and current trends in computing*. RSCTC 2000. Lecture Notes in Computer Science,vol 2005. Springer Berlin/Heidelberg, pp. 378-385
Gurol, Y. and Atsan, N. 2006. "Entrepreneurial characteristics amongst university students," *Education and Training* (48:1), pp. 25-38.
Henebry, K. L. 1996. "Do cash flow variables improve the predictive accuracy of a Cox proportional hazards model for bank failure?," *The Quarterly Review of Economics and Finance* (36:3), pp. 395-409.
Harabi, N. 2003. *Determinants of firm growth: An empirical analysis from Morocco*, MPRA Paper. Switzerland: University of Applied Sciences.
Hastie, T., R. Tibshirani, J. Friedman and Franklin, J. 2005. "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer* (27:2), pp. 83-85.
Hunt, R. J. 1986. "Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability," *Journal of Dental Research* (65:2), pp. 128-130.
Ibrahim, B. A. and Goodwin, J. R. 1986. "Perceived causes of success in small business," *American Journal of Small Business* (11:2), pp. 41-50.
Huselid, M. A., Jackson, S. E. and Schuler, R. S. 1997. "Technical and strategic human resources management effectiveness as determinants of firm performance," *Academy of Management journal* (40:1), pp. 171-188.
Judge, W. Q. and Douglas, T. J. 1998. "Performance implications of incorporating natural environmental issues into the strategic planning process: an empirical assessment," *Journal of management Studies* (35:2), pp. 241-262.
Kim, H. S. and Sohn, S. Y. 2010. "Support vector machines for default prediction of SMEs based on technology credit," *European Journal of Operational Research* (201:3), pp. 838-846.
Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection," In: *International Joint Conference on Artificial Intelligence*, pp. 1137-1145.
Kosala, R. and Blockeel, H. 2000. "Web Mining Research: A Survey," *ACM SIGKDD Explorations Newsletter* (2:1).
Kruppa, J., A. Schwarz, G. Arminger and Ziegler, A. 2013. "Consumer credit risk: Individual probability estimates using machine learning." *Expert Systems with Applications* (40:13), pp. 5125-5131.
Lev, B., Petrovits, C. and Radhakrishnan, S. 2010. "Is doing good good for you? How corporate charitable contributions enhance revenue growth," *Strategic Management Journal* (31:2), pp. 182-200.
Li, Y., S. Arora, J. Youtie and Shapira, P. 2016. "Using web mining to explore Triple Helix influences on growth in small and mid-size firms." *Technovation*.
Liedholm, C. 2002. "Small firm dynamics: Evidence from Africa and Latin America," *Small Business Economics* (18:1-3), pp. 227-242.
Lin, S. H. and Ho, J. M. 2002. "Discovering informative content blocks from Web documents," In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 588-593.
Lippitt G. L. and Schmidt, W. H. 1967. *Crises in a developing organization,* Harvard Business Review.

Lussier, R. N. and Halabi, C. E. 2010. "A three-country comparison of the business success versus failure prediction model. Journal of Small Business Management" (48:3), pp. 360-377.
OECD 2004. *Small and Medium-Sized Enterprises In Turkey Issues And Policies Organization For Economic Co-Operation And Development*, OECD Press.
O'Farrell, P. N. and Hitchens, D. M. 1988. "Alternative theories of small-firm growth: a critical review," *Environment and Planning A* (20:10), pp. 1365-1383.
Ohlson, J. A. 1980. "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research,* pp. 109-131.
Olson, E. M., Slater, S. F. and Hult, G. T. M. 2005. "The performance implications of fit among business strategy, marketing organization structure, and strategic behavior," *Journal of marketing* (69:3), pp. 49-65.
OSM Data for Switzerland 2016. http://planet.osm.ch/
Ozgulbas, N. and Koyuncugil, A. S. 2012. "Risk Classification of SMEs by Early Warning Model Based on Data Mining," *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* (6:10), pp. 2649-2660.
Mazzarol, T., Volery, T., Doss, N. and Thein, V. 1999. "Factors influencing small business start-ups," *International Journal of Entrepreneurial Behaviour and Research* (5:2), pp. 48-130.
Michie, J. and Sheehan, M. 2005. "Business strategy, human resources, labour market flexibility and competitive advantage," *The International Journal of Human Resource Management* (16:3), pp. 445-464.
Mitchell, R. 2015. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc.
Morinaga, S., K. Yamanishi, K. Tateishi and Fukushima, T. 2002. "Mining product reputations on the web," In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 341-349.
Mueller, S. and Thomas, A. S. 2001. "Culture and entrepreneurial potential: A nine country study of locus of control and innovativeness," *Journal of Business Venturing* (16:1), pp. 51-75.
Neiberg, D., Elenius, K. and Laskowski, K. 2006. „Emotion recognition in spontaneous speech using GMMs," In: *Interspeech*, pp. 809-812.
Nguyen, V. P., Laisney, F. and Kaiser, U. 2004. "The performance of German firms in the business-related service sector: A dynamic analysis," *Journal of Business & Economic Statistics* (22), 274-295.
Nooteboom, B. 1988. "The Facts About Small Business and the Real Values of Its Life World: A Social Philosophical Interpretation of This Sector of the Modem Economy," *American journal of economics and sociology* (47:3), pp. 299-314.
Patel, K. B., Chauhan, J. A. and Patel, J. D. 2011. "Web Mining in E-Commerce: Pattern Discovery, Issues and Applications," *International Journal of P2P Network Trends and Technology* (1:3), pp. 40-45.
Pearson, J. N. and Ellram, L. M. 1995. "Supplier selection and evaluation in small versus large electronics firms," *Journal of Small Business Management* (33:4), p. 53.
Pompe, P. P. and Bilderbeek, J. 2005. „The prediction of bankruptcy of small-and medium-sized industrial firms," *Journal of Business venturing* (20:6), pp. 847-868.
Post, H. A. 1997. "Building a Strategy on Competencies," *Long range Planning* (30:5), pp. 733–740.
Potter, J. G. and Storey, D. J. 2007. *OECD framework for the evaluation of SME and entrepreneurship policies and programmes,* Publications de l'OCDE.
Quayle, M. 2003. "A study of supply chain management practice in UK industrial SMEs," *Supply Chain Management: An International Journal* (8:1), pp. 79-86.
Reinartz, W., Krafft, M. and Hoyer, W. D. 2004. „The customer relationship management process: Its measurement and impact on performance," Journal of marketing research (41:3), pp. 293-305.
Reynolds, P. D., Hay, M., Bygrave, W. D., Camp, S. M. and Autio, E. 2000. *Global Entrepreneurship Monitor,* 2000 Executive Report. Babson College, Kauffman Center for Entrepreneurial Leadership, and London Business School.
Richard, O. C. 2000. "Racial diversity, business strategy, and firm performance: A resource-based view," *Academy of management journal* (43:2), pp. 164-177.
Robertson, M., Collins, A., Madeira, N. and Slater, J. 2003. "Barriers to start-up and their effect on aspirant entrepreneurs," *Education and Training* (45:6), pp. 308-316.
Sharda, D. and Chawla, S.. "Web Content Mining Techniques-A Study," *International Journal of Innovative Research in Technology and Science (IJIRTS).*

Scott, M. and Bruce, R. 1987. "Five Stages of Growth in Small Business," *Long Range Planning*, pp. 45-52.
Sokolova, M. and Lapalme, G. 2009. "A systematic analysis of performance measures for classification tasks," *Information Processing & Management* (45:4), pp. 427-437.
Steinmetz, L. L. 1969. "Critical stages of small business growth: When they occur and how to survive them," *Business horizons* (12:1), pp. 29-36.
Stewart, W. H., Carland, J. C., Carland, J. W., Watson, W. W. and Sweo, R. 2003. "Entrepreneurial dispositions and goal orientations: A comparative exploration of United States and Russian entrepreneurs," *Journal of Small Business Management* (41:1), pp. 27-47.
Swierczek, F. W. and Ha, T. T. 2003. "Entrepreneurial orientation, uncertainty avoidance and firm performance: An analysis of Thai and Vietnamese SMEs," *International Journal of Entrepreneurship and Innovation* (4:1), pp. 46-58.
Swiss Federal Statistical Office 2016. http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html. *Accessed 3 October 2016*
Swiss Federal Statistical Office: STAT-TAB 2016. http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html. *Accessed 3 October 2016*
Swiss Federal Tax Administration 2016. https://www.estv.admin.ch/. *Accessed 15 January 2017*
The World Business Council for Sustainable Development 2007. *A business guide to development actors,* Switzerland: Atar Roto Presse SA.
Thorleuchter, D. and Van Den Poel, D. 2012. "Predicting e-commerce company success by mining the text of its publicly-accessible website," *Expert Systems with Applications* (39:17), pp. 13026-13034.
Walker, E. and Brown, A. 2004. "What success factors are important to small business owners?," *International Small Business Journal* (22:6), pp. 577-593.
Wang, C. K. and Ang, B. L. 2004. „Determinants of venture performance in Singapore," *Journal of Small Business Management* (42:2), pp. 347-363.
Weiss, S. M. and Kulikowski, C. A. 1991. *Computer Systems that Learn Morgan Kaufman Publishers,* San Mateo.
West, D. 2000. "Neural network credit scoring models," *Computers & Operations Research* (27:11), pp. 1131-1152.
Wiklund, J., Patzelt, H. and Shepherd, D. A. 2009. "Building an integrative model of small business growth," *Small Business Economics* (32:4), pp. 351-374.
Wong, B. K., Lai, V. S. and Lam, J. 2000. "A bibliography of neural network business applications research: 1994–1998," *Computers & Operations Research* (27:11), pp. 1045-1076.
Worthington, I. and Britton, C. 2009. *Business environment,* Pearson Education.
Yusuf, A. 1995. "Critical success factors for small business: Perceptions of South Pacific entrepreneurs," *Journal of Small Business Management* (33:1), pp. 68-73.
Zhang, G., Hu, M. Y., Patuwo, B. E. and Indro, D. C. 1999. „Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *European journal of operational research* (116:1), pp. 16-32.