

Contents

1	Introduction	1
1.1	Internet, Big Data and Economic Indicators	1
1.1.1	Evolution in Internet use and the WWW	1
1.1.2	Internet and Big Data	4
1.1.3	Economic indicators	8
1.1.4	Online-based Big Data to produce economic indicators .	10
1.2	Motivation	13
1.3	Hypotheses and objectives	15
1.4	Structure of the thesis	16
2	Big Data sources and methods for social and economic analyses	19
2.1	Introduction	20
2.2	Related work	23
2.3	Non-traditional sources of social and economic data	26
2.3.1	The Internet as basic means for generating socio-economic data	29
2.3.2	Urban and mobile sensors	35
2.4	Non-traditional methods for processing social and economic data	36
2.4.1	Methods for structuring data	38
2.4.2	Methods for modelling data	39
2.4.3	Methods for assessing models' performance and robustness	42
2.5	The data lifecycle	44

CONTENTS

2.6	A Big Data Architecture for nowcasting and forecasting social and economic changes	50
2.6.1	Data analysis layer	50
2.6.2	Governance layer	58
2.6.3	Persistence layer	59
2.7	Conclusions	59
3	Web data mining for monitoring business export orientation	63
3.1	Introduction	64
3.2	Theoretical background	67
3.2.1	Web data mining for science and economic indicators . .	67
3.2.2	Export-related indicators built from website features . .	69
3.2.3	Structural variables related to export orientation	72
3.3	Using web-based variables to infer firm export orientation . . .	74
3.3.1	The sample	74
3.3.2	Data analysis	76
3.3.3	The predictive models	79
3.4	Automating the retrieval of web-based variables	82
3.4.1	Architecture of the web data mining system for analyzing corporate websites	83
3.4.2	Construction and validation of automatic web-based variables	84
3.4.3	Predicting firm export orientation from automatic web-based variables	88
3.5	Conclusions	90
4	Monitoring e-commerce adoption from online data	93
4.1	Introduction	94
4.2	Related work	97
4.3	SAME: a system for detecting and monitoring e-commerce adoption	100
4.3.1	The capture module	100

4.3.2	The analysis module	102
4.3.3	The production module	104
4.4	Experimental results	105
4.4.1	Data	105
4.4.2	Results	107
4.5	Conclusions	116
5	Do corporate websites' changes reflect firms' survival?	121
5.1	Introduction	122
5.2	Theoretical background	125
5.2.1	Firms' survival analysis	125
5.2.2	Capturing firms' economic activities through web data .	127
5.3	Data and methodology	129
5.3.1	Data	129
5.3.2	Multi-period logistic regression	132
5.3.3	Survival analysis	134
5.4	Results	135
5.4.1	Descriptive statistics and group comparisons	136
5.4.2	Multi-period logistic regression models	138
5.4.3	Survival analysis	141
5.5	Conclusions	142
6	Conclusions	145
6.1	Main contributions	145
6.2	Implications	147
6.3	Limitations	150
6.4	Future work	151
	Bibliography	153