UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Visual exploration to support the identification of relevant attributes in time-varying multivariate data

Aurea Rossy Soriano Vargas

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Aurea Rossy Soriano Vargas

Visual exploration to support the identification of relevant attributes in time-varying multivariate data

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Maria Cristina Ferreira de Oliveira

USP – São Carlos
April 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S714v Soriano Vargas, Aurea Rossy
Visual exploration to support the identification
of relevant attributes in time-varying multivariate
data / Aurea Rossy Soriano Vargas; orientadora
Maria Cristina Ferreira de Oliveira. -- São Carlos,
2018.
146 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2018.

1. Visual Analytics. 2. Time-varying
Multivariate Data. 3. Visual Feature Selection. 4.
Feature Space Analysis. 5. Exploratory Data
Visualization. I. Ferreira de Oliveira, Maria
Cristina, orient. II. Título.

Aurea Rossy Soriano Vargas

**Visualização como apoio à identificação de atributos
relevantes em dados multidimensionais variantes no tempo**

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Doutora em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientadora: Profa. Dra. Maria Cristina Ferreira
de Oliveira

USP – São Carlos
Abril de 2018

*This work is dedicated to my husband Oscar Cuadros Linares, to my father Raul Soriano
Filinich, and to my mother, Dora Vargas Núñez de Soriano.*

ACKNOWLEDGEMENTS

This is the culmination of an important and long journey of my life, my Ph.D. This thesis is the outcome of not only my efforts but also the help and guidance of many people, who have contributed to this huge task.

Foremost, I wish to give my heartfelt thanks to a very special person, my husband Oscar Cuadros. Thanks for his unconditional love, his unflagging encouragement, patience, and continued support that made the culmination of this thesis possible. I greatly value his insightful comments, the stimulating discussions, that helped me to keep things in perspective, the sleepless nights we were working together before deadlines and for all the fun we have had on this journey.

I would like to express my special appreciation and my sincere gratitude to my advisor Prof. Maria Cristina Ferreira de Oliveira. Thanks (i) for the continuous support and patience during all my Ph.D. study and research, especially in the most critical periods, (ii) for her helpful and priceless advice, and (iii) for her insightful comments. Also, I would like to thank Prof. Bernd Hamann for receiving me as a visiting research scholar at the University of California, Davis. Thanks for his guidance, for supporting my research ideas, and for his insightful discussions about the research in the last months. Both of them, Prof. Cristina and Prof. Bernd, were and remain my best role models for a scientist and mentor. ¡Muchas gracias!

I would like to extend my sincerest gratitude to the people who mean a lot to me, my parents Raul Soriano Filinich and Dora Vargas Núñez de Soriano. Thanks for their unconditional love, their emotional support, for all of the sacrifices that you have made on my behalf. Also, I express my thanks to my sister Laura, my little heart Fabio and my grandaunt Marlene Núñez (Pipito).

I would also like to dedicate this thesis to the memory of my grandmother Mamá Nelly Núñez. It was hard to say goodbye from afar. Even though the Alzheimer's demons attacked her, she treasured the memory of her granddaughter Aurea. Many times she talked with me about her granddaughter(me) with so much love, and joy because she(I) would visit her, not knowing that it was me. My mind and my heart treasure the great memories we had. Rest in peace.

I would like to thank all my friends from Brazil, Spain, United States and Peru, especially the Reginas girls, the Pipocas girls, and the Chelas guys, who always support me, give me good advice and make my life colorful.

Finally, I acknowledge the financial support provided by the State of São Paulo Research Foundation (FAPESP) grants 12/24537-0 and 15/12831-0.

“It will hurt. It will take time. It will require dedication. It will require willpower. You will need to make healthy decisions. It requires sacrifice. You will need to push your body to its max. There will be temptation. But, I promise you, when you reach your goal, it’s worth it. ”

(Unknown)

ABSTRACT

SORIANO-VARGAS, A. R. **Visual exploration to support the identification of relevant attributes in time-varying multivariate data**. 2018. 146 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Ionospheric scintillation is a rapid variation in the amplitude and/or phase of radio signals traveling through the ionosphere. This spatial and time-varying phenomenon is of interest because its occurrence may affect the reception quality of satellite signals. Specialized receivers at strategic regions can track multiple variables related to the phenomenon, generating a database of historical observations on the regional behavior of ionospheric scintillation. The analysis of such data is very challenging, since it consists of time-varying measurements of many variables which are heterogeneous in nature and with possibly many missing values, recorded over extensive time periods. There is a need to introduce alternative intuitive strategies that contribute to experts acquiring further knowledge from the ionospheric scintillation data.

Such challenges motivated a study on the applicability of visualization techniques to support tasks of identification of relevant attributes in the study of the behavior of phenomena described by multiple time-varying variables, of which the ionospheric scintillation is a good example. In particular, this thesis introduces a visual analytics framework, named *TV-MV Analytics*, that supports exploratory tasks on time-varying multivariate data and was developed following the requirements of experts on ionospheric scintillation from the Faculty of Science and Technology of UNESP at Presidente Prudente, Brazil. *TV-MV Analytics* provides an interactive visual exploration loop to analysts inspecting the behavior of multiple variables at different temporal scales, through temporal representations associated with clustering and multidimensional projection techniques. Analysts can also assess how different feature sub-spaces contribute to characterizing a certain behavior, where they may direct the analysis process and include their domain knowledge in the exploratory analysis.

We also illustrate the application of *TV-MV Analytics* on multivariate time-varying data sets from three alternative application domains. Experimental results indicate the proposed solutions show good potential on assisting time-varying multivariate data mining tasks, since it reduces the effort required from experts to gain deeper insight into the historical behavior of the variables describing a phenomenon or domain.

Keywords: Visual Analytics, Time-varying Multivariate Data, Visual Feature Selection, Feature Space Analysis, Exploratory Data Visualization.

RESUMO

SORIANO-VARGAS, A. R. **Visualização como apoio à identificação de atributos relevantes em dados multidimensionais variantes no tempo**. 2018. 146 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

A cintilação ionosférica é uma variação rápida na amplitude e/ou na fase dos sinais de rádio que viajam através da ionosfera. Este fenômeno espacial e variante no tempo é de grande interesse, pois pode afetar a qualidade de recepção dos sinais de satélite. Receptores especializados em regiões estratégicas podem rastrear múltiplas variáveis relacionadas ao fenômeno, gerando um banco de dados de observações históricas sobre o comportamento regional da cintilação. O estudo do comportamento da cintilação é desafiador, uma vez que requer a análise extensiva de dados multivariados e variantes no tempo, coletados por longos períodos. Medições são registradas continuamente, e são de natureza heterogênea, compreendendo múltiplas variáveis de diferentes categorias e possivelmente com muitos valores faltantes. Portanto, existe a necessidade de introduzir estratégias alternativas, eficientes e intuitivas, que contribuam para a aquisição de conhecimento, a partir dos dados, por especialistas que estudam a cintilação ionosférica.

Tais desafios motivaram o estudo da aplicabilidade de técnicas de visualização para apoiar tarefas de identificação de atributos relevantes no estudo do comportamento de fenômenos ou domínios que envolvem múltiplas variáveis, como a cintilação. Em particular, esta tese introduz um arcabouço visual, o qual foi denominado *TV-MV Analytics*, que apoia tarefas de análise exploratória sobre dados multivariados e variáveis no tempo, inspirado em requisitos de especialistas no estudo da cintilação, vinculados à Faculdade de Ciências e Tecnologia da UNESP de Presidente Prudente, Brasil. O *TV-MV Analytics* fornece aos analistas um ciclo de interativo de exploração que apoia a inspeção do comportamento temporal de múltiplas variáveis, em diferentes escalas temporais, por meio de representações visuais temporais associadas a técnicas de agrupamento e de projeção multidimensional. Também permite avaliar como diferentes sub-espacos de atributos caracterizam um determinado comportamento, podendo direcionar o processo de análise e inserir seu conhecimento do domínio no processo de análise exploratória.

As funcionalidades do *TV-MV Analytics* também são ilustradas em dados variantes no tempo oriundos de outros três domínios de aplicação. Os resultados experimentais indicaram que as soluções propostas têm bom potencial em tarefas de mineração de dados multivariados e variantes no tempo, uma vez que reduz o esforço e contribui para os especialistas obterem informações detalhadas sobre o comportamento histórico das variáveis que descrevem um determinado fenômeno ou domínio.

Palavras-chave: Análise Visual, Dados Multivariados e Variantes no Tempo, Seleção Visual de Atributos, Análise de Espaço de Atributos, Visualização Exploratória de Dados.

LIST OF FIGURES

Figure 1 – InterRing from a hierarchy of 42 variables (YANG <i>et al.</i> , 2003b).	32
Figure 2 – Dimension Projection Tree/Matrix (YUAN <i>et al.</i> , 2013).	34
Figure 3 – Model of double visual analysis, developed on the two spaces: data instances and variables (TURKAY; FILZMOSER; HAUSER, 2011).	34
Figure 4 – Visualization of 1,430 stock prices along 52 weeks (HOCHHEISER; SHNEIDERMAN, 2004).	38
Figure 5 – TimeSearcher (HOCHHEISER; SHNEIDERMAN, 2004).	39
Figure 6 – CircleView (KEIM; SCHNEIDEWIND; SIPS, 2004).	40
Figure 7 – Cluster and Calendar based Visualization of Time Series Data (WIJK; SELOW, 1999).	40
Figure 8 – VizTree (LIN <i>et al.</i> , 2004).	41
Figure 9 – Interactive Visual Analysis (IVA) cycle (TURKAY <i>et al.</i> , 2011).	42
Figure 10 – Bitmap relative to four time series (KUMAR <i>et al.</i> , 2005).	42
Figure 11 – System by Steiger <i>et al.</i> (2014).	43
Figure 12 – Overview of the pipeline of the Correlated Multiples algorithm (LIU <i>et al.</i> , 2015).	43
Figure 13 – Interface of the TimeSpiral System.	44
Figure 14 – VIMTEX system (DASGUPTA; KOSARA; GOSINK, 2015).	45
Figure 15 – Falcon system. The main window consists of interactive visualizations that allow users to explore temporal and statistical patterns in multivariate time series. Two additional analysis windows are available to provide more detailed views of segments derived from the entire time series of a particular variable (STEED <i>et al.</i> , 2017).	46
Figure 16 – Player Attribute Heatmap (PAH) visualization of the positions of players during a match (MACHADO <i>et al.</i> , 2017).	46
Figure 17 – Frequency of occurrence of ionospheric scintillation according to geographic location (KINTNER; HUMPHREYS; HINKS, 2009).	51
Figure 18 – Stations of the CIGALA/CALIBRA Network (VANI <i>et al.</i> , 2013).	51
Figure 19 – Examples of point-based visualizations in the ISMR Query Tool (VANI, 2014) (UNESP)	58
Figure 20 – Grid-based visualization: The s4 index relative to all GPS satellites from station PRU1 (VANI, 2014).	58
Figure 21 – Pierce points visualization (VANI, 2014).	59

Figure 22 – Visualizing the behavior of 6 variables on the horizon charts (VANI, 2014). .	59
Figure 23 – Visualization using the SAX technique (VANI, 2014).	60
Figure 24 – S4 index variation in a West African equatorial region (ACKAH; OBROU; GROVES, 2011).	60
Figure 25 – Map of the regions for which crime statistics are recorded in State of São Paulo, Brazil.	61
Figure 26 – Clusters of segments of 550 stocks of software and banking sector (ZIEGLER <i>et al.</i> , 2010).	65
Figure 27 – Overview of <i>TV-MV Analytics</i>	69
Figure 28 – Software Architecture.	71
Figure 29 – General database model.	71
Figure 30 – Query window: the initial window of <i>TV-MV Analytics</i> is divided into four sections for selecting: (A) a subset of target variables, (B) the temporal range, aggregation function and analysis time unit, (C) the threshold options and source, and (D) matrix moments.	74
Figure 31 – Colormap used to characterize each resulting cluster.	80
Figure 32 – Five strategies for variable space selection.	81
Figure 33 – Time matrix visualization of variable <i>phi0111</i> , depicting daily data.	83
Figure 34 – Time matrix visualization of the variable <i>phi0111</i> , depicting monthly data. .	84
Figure 35 – Heated-Object colormap (LEVKOWITZ, 1997).	84
Figure 36 – Example of prediction task using Isotonic Regression. An isotonic function is created for each colum. The series values in the next-row are predicted using that function based on historical data.	85
Figure 37 – Schematic visualization of the hierarchical time matrices. Circles represent feature vector of time matrices, which are the input data for the clustering tree. The root contains the medoids of each input cluster, each level contains submedoids, and leaves store each feature vector point.	86
Figure 38 – Time Circular Diagram of a variable in the air quality data set, showing the correlations of 1-day periods from March 10-31, 2004. The sub-periods are placed clockwise, where the initial day (March 10) is placed at the middle top.	88
Figure 39 – Neighborhood Preservation curves depicting precision of similarity maps of the 70 ionospheric scintillation variables obtained with four distinct projection techniques. Clearly, IDMAP yields precision rates superior to LSP and comparable to those obtained by classic approaches such as Fastmap and PCA.	90
Figure 40 – Schematic view of the hierarchical similarity map of variables. A KD tree spatial representation (SAMET, 1990) of the 2D embedded variable points. The root contains projected variable points of all space, and the leaf nodes individual variables	91

Figure 41 – Correlation matrix of Pearson coefficients computed from aggregated scintillation variables. Matrix rows and columns are sorted according to the clusters of variables, indicated by the colored lines along matrix borders.	92
Figure 42 – Normalized time series graphs and timelines depicting the corresponding time-matrix values.	93
Figure 43 – IDMAP similarity map and small multiple time matrices view of 70 variables describing ionospheric scintillation data instances.	97
Figure 44 – Variable Correlation Matrix (Pearson coefficient) for a specific subset of variables assigned to three distinct clusters (C4-lilac, C5-red and C6-brown).	98
Figure 45 – IDMAP similarity map view of the same 70 variables depicted in Figure 43.A.	99
Figure 46 – Similarity maps of 8,868 data instances described by three variable spaces.	100
Figure 47 – A three-day sample of the normalized time series of aggregated values (maximum over 1h intervals) for variables <i>avg_cn0_11</i> and <i>S4_corrected</i> , where a daily seasonal pattern can be observed.	101
Figure 48 – Cluster models of the 70 variables in terms of their size and composition (indicated by colors), obtained from data instances aggregated at different time intervals: three hours, two hours, one hour, 30 minutes, 20 minutes, 10 minutes and five minutes. Cluster patterns differ over the different levels of granularity, although a few clusters remain quite stable.	102
Figure 49 – Actual and predicted values (January 31, 00:00h-10:00h) of variables (a) <i>S4_corrected</i> ; (b) <i>Si_11</i> . Prediction of <i>S4_corrected</i> shows a correlation of 0.9731 with actual values, whereas for <i>Si_11</i> correlation is 0.8863.	102
Figure 50 – IDMAP similarity map and Small Multiple Time Matrices view of values of index <i>S4_corrected</i> recorded at the 32 GPS satellites.	103
Figure 51 – Occurrences of drug possession, drug seizure, drug traffic and illegal weapon possession in three regions.	105
Figure 52 – IDMAP similarity map view of the same indicators depicted in Figure 51.A, in which circle colors reflect an alternative (user-adjusted) cluster model (4 clusters).	106
Figure 53 – Time Matrices of <i>total number of inquiries</i> by region associated to the Map of São Paulo, Brazil.	107
Figure 54 – Map of the state of São Paulo conveying the clustering model for each crime indicator.	108
Figure 55 – Actual and predicted values of <i>imprisonments</i> at the state of São Paulo in 2016. The correlation between actual and predicted is 0.76.	109
Figure 56 – Time Matrices of six crime indicators.	110
Figure 57 – Time series graphs of six crime indicators.	111
Figure 58 – Visualization of 14 indicators for the São Paulo State.	112

Figure 59 – IDMAP similarity map view of the same 14 indicators depicted in Figure 58, now considering an user-adjusted cluster model (5 clusters).	112
Figure 60 – Time Matrices of air quality data considering the entire period March 2004 - April2005.	116
Figure 61 – Correlation Matrix of original and aggregated values measured for variables: C_6H_6 , plus CO , NO_2 and NO_x (reference and MOX).	118
Figure 62 – Line graphs depicting the normalized series of measurements of $ST_C_6H_6$ and MOX_CO	118
Figure 63 – Time circular diagram of $ST_C_6H_6$ and MOX_CO , considering values from May 15 to May 22, 2004 and 1-day sub-periods.	119
Figure 64 – Time Matrices and similarity map for the 12 US companies.	121
Figure 65 – Three intervals of historical data of variable <i>close</i> for IBM.	122
Figure 66 – Hierarchical time matrices of the variable <i>close</i> for 438 companies.	122
Figure 67 – First and second hierarchy levels of the hierarchical time matrix structure depicting the technological companies.	124
Figure 68 – The red group: high values occur mostly at the beginning of the observed period, followed by low values.	124
Figure 69 – The blue group: low values occur at the beginning and high values occur at the end of the period.	125
Figure 70 – The yellow group: characterized by a smooth variation in intensities.	125
Figure 71 – The lilac group: a high percentage of missing data and high values.	126
Figure 72 – The green group: also characterized by a high percentage of missing data, but a different pattern of intensity in the values.	127
Figure 73 – Basic example of an XML configuration file structure.	146

LIST OF TABLES

Table 1 – Related work on visualization and analysis of variable spaces in multidimensional data.	38
Table 2 – Features of related work on visualization and analysis of multivariate time series.	47
Table 3 – A time-varying multivariate data set, where each row describes a time-stamped data instance and each column describes a particular variable. Data instances describe year-average physiological measures of a Pakistani boy from 1997 to 2002 (AZIZ <i>et al.</i> , 2012).	49
Table 4 – Identifiers considered in the ionospheric scintillation data (Septentrio Satellite Navigation, 2013).	52
Table 5 – Diverse scintillation indices that describe the behavior of ionospheric scintillation data (Septentrio Satellite Navigation, 2013).	53
Table 6 – S4 indices related to the behavior of ionospheric scintillation of amplitude (Septentrio Satellite Navigation, 2013).	54
Table 7 – Categorization of the S4 index according to different authors.	55
Table 8 – Sigma-phi indices related to the behavior of ionospheric scintillation of phase (Septentrio Satellite Navigation, 2013).	56
Table 9 – Categorization of sigma-phi index according to different authors.	56
Table 10 – Variables related to the synchronization time and GPS time of ionospheric scintillation (Septentrio Satellite Navigation, 2013).	57
Table 11 – Variables related to the spatial orientation of ionospheric scintillation (Septentrio Satellite Navigation, 2013).	57
Table 12 – Crime indicators	62
Table 13 – Variables of UCI air quality data set (VITO <i>et al.</i> , 2008).	63
Table 14 – Data sets analyzed.	66
Table 15 – Different choices of feature vectors to represent the internal time matrices. # represents the size of the resulting feature vectors.	76
Table 16 – Statistical Moments (Color Moments).	77
Table 17 – Hu moments (HU, 1962).	77
Table 18 – High-level analysis questions of ionospheric scintillation data identified in collaboration with domain experts.	96
Table 19 – Classification accuracy and regression correlation of data instances described by three distinct variable subspaces.	101
Table 20 – Analysis questions of crime statistics of São Paulo State, Brazil.	104

Table 21 – Silhouette coefficient(SC) of different clustering models of four indicators related to occurrences of drug (3) and illegal weapon possession in the crime statistics data.	106
Table 22 – Correlations between actual and predicted values of indicator <i>imprisonments</i> using historical data from January 31, 2002 to December 31, 2015. Most correlations are positive, which indicates that it is possible to predict the values of this indicator. However, the Santos indicator has a negative correlation, which would indicate the presence of other factors.	109
Table 23 – Regression correlation of data instances described by the distinct indicator subspaces. # is the space number, # ind. the number of selected indicators and corr. the correlation.	114
Table 24 – Regression correlation of data instances described by the distinct indicator subspaces. # is the space number, # ind. the number of selected indicators and corr. the correlation.	115
Table 25 – Most correlated variables with each polluting gas.	117
Table 26 – Correlations between measurements of different pollutants obtained from the reference device (spectrometer analyzer) as computed by Vito <i>et al.</i> (2008) and as obtained in our framework. Our results, obtained through interacting with the views, confirm that aggregated data instances suffice to characterize the target variables.	117
Table 27 – Analysis questions of stock market data.	119
Table 28 – US biggest technological companies in 2016 (according to Forbes) listed on Nasdaq Stock Market.	120
Table 29 – Prediction results of variable <i>close</i> for IBM in 2016, using different intervals of historical data: 31, 15 and 5 years.	121
Table 30 – Questions defined by Dasgupta et al. (2015) (DASGUPTA; KOSARA; GOSINK, 2015)	130
Table 31 – Variables that describe the behavior of ionospheric scintillation (Septentrio Satellite Navigation, 2013) (Part 1). The gray rows refer to identifiers.	143
Table 32 – Variables that describe the behavior of ionospheric scintillation (Septentrio Satellite Navigation, 2013) (Part 2). The gray rows refer to identifiers.	144

LIST OF ABBREVIATIONS AND ACRONYMS

C_6H_6	Benzene
CO	Carbon Monoxide
$NMHC$	Non-Methane Hydrocarbons
NO_2	Nitrogen dioxide
NO_x	Nitrogen Oxide
CIGALA	Concept for Ionospheric Scintillation Mitigation for Professional GNSS in Latin America
GNSS	Global Navigation Satellite Systems
GPS	Global Positioning System
LSP	Least Square Projection
MOX	Metal Oxides
NP	Neighborhood Preservation
OVM	Orthogonal Variant Moments
PAH	Player Attribute Heatmap
SC	Silhouette Coefficient
TV-MV Analytics	Visual Time-Varying Multivariate Data Analytics framework
VICG	Visualization, Imaging and Computer Graphics laboratory

CONTENTS

1	INTRODUCTION	25
1.1	Context	25
1.2	Motivation	26
1.3	Hypothesis and Goals	27
1.4	Contribution	28
1.5	Organization	29
2	RELATED WORK	31
2.1	Initial Remarks	31
2.2	Visualization and Analysis of Variable Spaces	32
2.3	Visualization and Analysis of Variables in Connection with Specific Mining Tasks	36
2.4	Exploratory Visualization of Multivariate Time Series	37
2.5	Chapter Notes	47
3	TIME-VARYING MULTIVARIATE DATA SETS	49
3.1	Initial Remarks	49
3.2	Ionospheric Scintillation Data	50
3.2.1	<i>Indicators and Measurements of Ionospheric Scintillation</i>	51
3.2.2	<i>Previous Work on Visualizing Ionospheric Scintillation Data</i>	57
3.3	Other Data Sets	61
3.3.1	<i>Crime Statistics for the State of São Paulo, Brazil</i>	61
3.3.2	<i>Air Quality Data at an Italian City</i>	62
3.3.3	<i>Stock Market Data</i>	64
3.4	Chapter Notes	66
4	THE VISUAL ANALYTICS FRAMEWORK	67
4.1	Initial Remarks	67
4.2	Overview of <i>TV-MV Analytics</i> Framework	68
4.3	Software Architecture	70
4.4	Query Window	72
4.5	Data Processing	73
4.5.1	<i>Time Matrix Representation</i>	74
4.5.2	<i>Feature Moments</i>	75

4.5.3	<i>Data Analytics</i>	79
4.6	<i>Visualization of Variables</i>	82
4.6.1	<i>Small Multiples Time Matrix Visualizations</i>	83
4.6.2	<i>Time Circular Diagram</i>	88
4.6.3	<i>Similarity Map of Variables</i>	89
4.6.4	<i>Variable Correlation Matrix</i>	92
4.7	<i>Visualization of Data Instances</i>	93
4.8	<i>Chapter Notes</i>	94
5	EXPERIMENTAL RESULTS	95
5.1	<i>Initial Remarks</i>	95
5.2	<i>Visual Analytics of Ionospheric Scintillation Data</i>	95
5.2.1	<i>Identifying Variables with Similar Temporal Behavior (Q1)</i>	96
5.2.2	<i>Identifying Relations between S4 Indices and Other Variables (Q2) and Relevant Variables to Scintillation (Q4)</i>	99
5.2.3	<i>Analysis at Multiple Temporal Granularities (Q3)</i>	101
5.2.4	<i>Variable Prediction (Q5)</i>	102
5.2.5	<i>Comparing Data Instances from Multiple Satellites (Q6)</i>	103
5.3	<i>Exploring Data from Other Domains</i>	104
5.3.1	<i>Crime Statistics of São Paulo State, Brazil</i>	104
5.3.2	<i>Air Quality Data of an Italian City</i>	116
5.3.3	<i>US Stock Market Data</i>	119
5.4	<i>Chapter Notes</i>	127
6	CONCLUSIONS	129
6.1	<i>Contribution</i>	129
6.2	<i>Limitations</i>	131
6.3	<i>Future work</i>	132
	BIBLIOGRAPHY	133
APPENDIX A	IONOSPHERIC SCINTILLATION VARIABLES	143
APPENDIX B	XML CONFIGURATION FILE	145

INTRODUCTION

1.1 Context

The rapid advancement of software and hardware technology allows users to expand their production, service, communication and research capacity. Consequently, little effort is required to accumulate immense amounts of multivariate data. However, analyzing such data requires handling the problem known as *dimensionality curse* (KORN; PAGEL; FALOUTSOS, 2001), as well as gaining insight on how data evolves over time (JÄCKLE *et al.*, 2016).

Time-varying multivariate data is a sub-category of data streams, in which data instances are described by multiple time-stamped variables recorded sequentially (YAZDANBAKHS; DICK, 2017). This type of data plays an important role in many domains, such as environmental monitoring, monitoring of natural phenomena, financial markets, population statistics and others (BERNARD *et al.*, 2012). Data analysis in multiple domains require the ability to thoroughly explore the variables, with the objective of identifying patterns, analyzing their behavior (STEED *et al.*, 2017) and making sense of long multivariate time series (BERNARD *et al.*, 2016).

Exploring and organizing time-varying multivariate data for representation to expert analysts or regular users is particularly challenging (MARTIN; QUACH, 2016). Typically, this data contains hundreds, thousands or even millions of instances; and sometimes analysis is conducted with limited prior knowledge, if any. Another common user task in this scenario is to identify subsets of variables responsible for characterizing a particular behavior; or the relevant variables to explain a phenomenon under study. Handling a large number variables impacts both the complexity of the analysis tasks, and the computational cost of the exploratory visualization and data mining methods (HAN, 2005; GHEYAS; SMITH, 2010). Many methods have been proposed to reduce the number of variables, but they often have a “black box” nature, which makes it impossible for the analysts to inspect the entire process or incorporate their expert knowledge and experience (KENNEDY; TURLEY, 2011).

Users are mainly interested in consuming this data, which means to present, to discover and to enjoy, generating or verifying hypothesis (MUNZNER, 2014). Nevertheless, due to the number of variables and instances, it becomes difficult to explore the interplay between the multiple variables recorded and/or data instances. They also need to explore the data considering different temporal scales, where scale refers to a certain aggregation for analysis (e.g., minute, hour, days, month, year). Hence, a flexible navigation through data represented at different temporal aggregations is necessary, while preserving the context (ZHANG *et al.*, 2016b).

Approaches in Visual Analytics, also known as Visual Data Mining, (OLIVEIRA; LEV-KOWITZ, 2003; HABIBI; SHIRKHODAIE, 2012; SUN *et al.*, 2013) – which explore the synergy between Data Mining from machine learning and data representations from Information Visualization – have been proposed to support processes that require extracting information from data. They have been successfully applied to several problem domains (INSELBERG, 2009; ENGEL *et al.*, 2012). The key is the interaction with graphical representations of data to favor understanding its behavior. Interaction allows integrating the human capability of recognizing visual patterns (inference capability) and users' knowledge into (automatic or semi-automatic) data analysis processes (SUN *et al.*, 2013). Thus, visual analytics approaches have become indispensable means to give analysts an adequate mental model to understand complex data, favoring the perception of global characteristics and the identification of local behavior through user-driven exploration (KEIM *et al.*, 2008). We also assume that appropriate visualization techniques favour a better understanding of how the multiple variables are related and how they relate to the instances recorded over time.

1.2 Motivation

This project resulted from an interdisciplinary collaboration with researchers from the Department of Cartography of Universidade Estadual Paulista (UNESP) at Presidente Prudente, who study the effects of ionospheric scintillation, a phenomenon that affects the reception quality of satellite signals. Specifically, it affects measurements obtained by Global Navigation Satellite Systems (GNSS) such as the Global Positioning System (GPS). The ionospheric scintillation is a change in the amplitude and/or phase of a radio signal, due to its propagation through regions of the ionosphere with irregularities in the electron density, causing amplitude attenuations and phase shifts in the GNSS signals (VANI, 2014). Its occurrence affects the reception quality of satellite signals, whereby the received information shows spurious and/or missing data.

As part of the Concept for Ionospheric Scintillation Mitigation for Professional GNSS in Latin America (CIGALA) project funded by the European Union, GNSS receivers were installed in Brazil to monitoring the phenomenon (ALFONSI *et al.*, 2011), and a total of 70 variables are measured or computed every minute for each satellite tracked. Scintillation data consists of multivariate time series of measurements and indices recorded in an application-specific

database.

There is great interest in analyzing the effects of scintillation and understanding its causes in order to gather elements to address its effects. The CIGALA database provides an important source of information, as it registers an expressive volume of measurements at several spatial locations. Although tools have already been developed to investigate this data (VANI *et al.*, 2013; VANI, 2014), additional solutions targeted at the complex nature of both the data and the questions posed by the experts are still needed.

1.3 Hypothesis and Goals

Our hypothesis is that exploratory visualization can be successfully applied to the CIGALA database to assist experts in the ionospheric scintillation domain application to answer relevant questions about the data. Specifically, they need to identify variables and corresponding behaviors that determine and explain the occurrence of high scintillation indices. Knowledge about the behavior of those variables contributes to better understanding the problem and enabling more effective monitoring of the phenomenon, possibly reducing the number of variables to be monitored and indices to be computed, and also leading to the formulation of predictive models.

The requirements drawn from the experts laid the groundwork for the framework developed: the visualizations should (i) help analysts to inspect and explore scintillation data at different temporal scales; (ii) provide resources to investigate the behavior of both individual variables and groups of variables; and (iii) include functionalities for exploring alternative sub-spaces of variables identified as representative to characterize the behavior of groups of scintillation measurements. As an additional validation effort, the proposed analytics strategies have also been applied to data collected from other domains (crime statistics, air quality and stock market) where large time-varying multivariate data arise and must be understood.

The main goal of this research is to develop and validate strategies, compiled in a visual framework, to support user-driven explorations of time-varying multivariate data. The focus is a visual framework conveying the interplay between multiple variables describing data instances recorder over time; the identification of variables that determine the behavior of a time-varying phenomenon; and temporal correlations between variables. Our Visual Time-Varying Multivariate Data Analytics framework (TV-MV Analytics) integrates data mining and visualization techniques, based on feature extraction, clustering and multiple interactive visualizations focused on data variables and instances. Other specific objectives are listed below:

1. Create visual representations that convey relevant temporal patterns.
2. Allow including expert knowledge in identifying representative variables for characterizing a particular behavior or variable.

3. Evaluate how different variable spaces suffice to characterizing the behavior of certain variables, using standard classification and regression algorithms.
4. Support interactive exploration of the similarities between groups of data instances and groups of data variables.

1.4 Contribution

Our devised integrative approach incorporated into the *TV-MV Analytics* system addresses the aforementioned requirements by combining multiple visualization techniques and mining algorithms. The system, developed as a proof-of-concept:

1. Provides data analysts with a visual exploration loop that supports inspecting, at different temporal scales, series of recorded multivariate values.
2. Derives a representation of the monitored variables that conveys their behavior in detail, at multiple user-defined aggregation levels.
3. Provides hierarchical visualizations that convey the temporal behavior of individual variables and also of groups of related variables.
4. Reveals how the multiple variables are related, by means of a small multiples visualization associated with clustering and multidimensional projection techniques.
5. Allows interactive exploration of the behavior of multiple feature spaces describing the data instances.
6. Allows analysts to assess how different feature sub-spaces characterize a certain behavior.

The initial framework has been devised specifically for the ionospheric scintillation problem, discussed in Chapter 3, and validated in this particular domain. A generalization was then abstracted from the integrated mining and visualization solution, which is applicable to other domains (also discussed in Chapter 3) with similar characteristics, i.e. that may benefit from exploratory analysis of time-varying multivariate data. As a complementary validation effort, exploration tasks were also conducted on these additional data sets. We believe this effort can assist data analysts who rely on visual data mining to better understand the problem scenario and requirements from the data. This is potentially useful in situations where it is not possible to accurately formulate the problem in order to apply conventional machine learning strategies.

1.5 Organization

This thesis is organized as follows.

- In Chapter 2 a literature survey is presented considering related work on feature selection of variables with support of visual representations, usage of visualizations to explore feature spaces describing multidimensional data, as well as recent work on visualization applied to assist the analysis of time-varying multivariate data.
- Chapter 3 details the time-varying multivariate data sets considered in our experimental validation and data processing issues. The main goal of our contribution was the analysis of ionospheric scintillation data from two Brazilian monitoring stations. But this solution can be useful in other scenarios, such as data originated from the US stock market, air quality measurements of an Italian city, and crime statistics of the State of São Paulo, Brazil.
- Chapter 4 introduces and describes the *TV-MV Analytics* framework, including a pipeline with the multiple interactive visualizations of both data variables and data instances.
- We illustrate alternative applications of the framework in four case studies in Chapter 5, related to the considered domains. Case studies address plausible data analysis scenarios and assess the framework's capability to inform the relevant variables that explain an observed behavior.
- Finally, Chapter 6 presents the conclusions, limitations and discusses possible future research.

RELATED WORK

2.1 Initial Remarks

Time-varying multivariate data or multivariate time series play an important role in many application domains ([JÄCKLE *et al.*, 2016](#)), where specialists face a growth in both the number of data instances and the number of describing variables. Consequently, multiple challenges are posed to visualization designers ([LIVINGSTON; DECKER; AI, 2013](#)), as conventional graphical representations quickly become overcrowded.

There is no consistent definition in the literature regarding the minimum number of variables that characterize a high-dimensional data set. [Assent \(2012\)](#) states that in some scenarios data described by only ten variables is considered high dimensional, while in others, especially in bioinformatics or image processing, data can be described by hundreds or thousands of variables. According to [Yang and Moody \(2000\)](#) and [Razente *et al.* \(2004\)](#), dimensionality reduction techniques and visualization strategies are closely related, offering great potential assistance in discovering hidden information and finding representative variables. There are previous contributions in the literature that integrate visualization and data exploration tasks to identify relations between variables and data instances and assist investigation of the variable spaces that describe multivariate data.

In this chapter, we briefly introduce techniques for analysis of variable spaces that rely on visual representations either to support general data exploration tasks focused on investigating the variable spaces (Section [2.2](#)), or to assist specific mining tasks such as clustering or classification (Section [2.3](#)), seeking to understand how variables affect the observed relationships between data instances. Due to the time-varying multivariate nature of the ionospheric scintillation data, we also review previous work on the analysis of multivariate time series in general (Section [2.4](#)).

2.2 Visualization and Analysis of Variable Spaces

Papers reviewed in this section deal with the analysis of variable spaces supported by visual representations to compare or assess the role of different variable spaces, for purposes of exploration, understanding, or mining tasks.

Strategies for user-driven exploration of the variable spaces defining a high-dimensional data set were proposed by [Yang *et al.* \(2003a\)](#), [Yang *et al.* \(2003b\)](#) and [Yang *et al.* \(2007\)](#). The first strategy, called Visual Hierarchical Dimension Reduction (VHDR) ([YANG *et al.*, 2003b](#)), allows generating lower-dimensional representation -spaces. Variables are grouped into a hierarchy, which is displayed in a radial visualization called InterRing, shown in Figure 1. Users can interactively explore and modify this hierarchy and select interesting clusters. A representative variable is determined for each selected cluster and the high-dimensional data is mapped into the resulting subspace defined by them, and displayed with traditional visualization techniques, such as Parallel Coordinates and dispersion graph arrays.

Figure 1 – InterRing from a hierarchy of 42 variables ([YANG *et al.*, 2003b](#)).

In a similar context [Yang *et al.* \(2003a\)](#) propose the Dimension Ordering, Spacing, and Filtering Approach (DOSFA), for automatic and interactive ordering, spacing, and filtering of variables, by grouping variables into a variable hierarchy derived from their similarities, to explore high dimensional data sets, reducing the complexity of those tasks.

A hierarchy of the variables is built employing a clustering algorithm and users can interact with the InterRing visualization to explore and modify this hierarchy. In the variable ordering task, the authors propose orderings based either on similarity or on importance. The similarity-oriented ordering considers an in-depth traversal in the hierarchy, not considering internal nodes; in the importance-oriented ordering the user interacts to determine which is the most important feature to be considered, for example, variance. The spacing task reports

on the variable relationships varying the space between adjacent axes or angles, positioning strongly related variables close, e.g., in Parallel Coordinates. Finally, the filtering task criterion combines similarity and importance information. From the hierarchy root, the algorithm selects the variables to be displayed, ignoring the less important ones.

The Value and Relation (VaR) technique (YANG *et al.*, 2007) aims to facilitate data exploration, visualization, and variable selection, helping users to understand relationships between variables and between instances. Data clusters and outliers can be detected in spaces defined by subsets of variables. The variables are projected with Multidimensional Scaling (MDS), where each variable is represented by a glyph. The glyphs are dense pixel displays of the corresponding data values mapped into a spiral arrangement of pixels, which can reveal patterns in the corresponding variable. Navigation and variable selection tools are provided to reduce overlap and enable interactive exploration.

There are also approaches aimed at **comparing multiple variable spaces**. In this context, Schreck, Fellner and Keim (2008) consider the problem of analyzing and comparing different variable spaces describing multimedia data. They propose an unsupervised estimator of the discrimination power of a variable space, assuming that it can be estimated from the uniformity of the histogram of the distances between the data clusters described in that space. The Self-Organizing Map (SOM) algorithm is employed to generate the clusters associated with the different spaces. In addition to mining and visualization tasks, the representation issue is also relevant to Content-Based Image Retrieval (CBIR). Many CBIR systems use the concept of Relevance Feedback (RF), in which users, based on their knowledge, train the system informing the perceived relevance of the hits returned from a query made. In this scenario, Ng and Martin (2004) propose a system that allows visualizing the global RF scores associated with the images returned from a query, as well as the values of their describing variables. They introduce a method to automatically compute the importance of a particular variable, given the relevance information. Thus, an analyst can assess the importance of the different variables in the query result.

The approach by Yuan *et al.* (2013) relies on projections. Based on the divide-and-conquer strategy, data sets are split or reduced into subsets of variables (called dimensions in the paper) and/or data instances, using a platform called Dimension Projection Matrix/Tree. In the Dimension Projection Matrix (an extension of a scatterplot matrix) each matrix row or column represents a group of variables, and each cell depicts a dimension projection (such as Principal Component Analysis (PCA) or Multidimensional Scaling (MDS)) of the corresponding variables. In the Dimension Projection Tree, each node is a projection graph or a projection matrix. Each child node contains subsets of variables or data instances from the parent node. Both functionalities, the Dimension Projection Matrix and the Dimension Projection Tree, can be built automatically or interactively by the user, so the hierarchy of data instances and variables can be explored at different levels of detail to identify correlations in data instances and variables.

Figure 2 illustrates an exploration using the Dimension Projection Tree/Matrix visualization. Each node contains one data item plot and one dimension plot. From an initial node (1), users can explore either in the data instance space (2, 3) or variable space (4, 5, 6), by first selecting a subset of data instances or variables, and then creating child nodes (3, 4, 5) or converting one single dimension project plot into a matrix representation (6).

Figure 2 – Dimension Projection Tree/Matrix (YUAN *et al.*, 2013).

In a similar approach considering techniques from multivariate statistical analysis, such as PCA or Linear Discriminant Analysis (LDA), Turkay, Filzmoser and Hauser (2011) propose a visualization model for interactive visual analysis of multivariate data focusing on both the variables and the data instances by selection mechanisms, presented in Figure 3. The interactions allow the joint and associated exploration of variable statistics and Multivariate Analysis Results (MVA), supporting the study of the structure of variables and distributions of data instances.

Figure 3 – Model of double visual analysis, developed on the two spaces: data instances and variables (TURKAY; FILZMOSER; HAUSER, 2011).

Visualization-assisted variable selection can also rely on **statistical techniques and**

measures. [May et al. \(2011\)](#) propose a method for semi-automatic refinement of results from variable selection algorithms, through a visual interface called SmartStripes. Given a multidimensional data matrix, it is possible to investigate the dependencies and inter-dependencies between different variable spaces and subsets of data instances in order to identify an optimal subset of variables, i.e., those relevant to a particular task and not redundant. The solution implemented treats numeric, ordinal, and nominal variables uniformly; which can be alternated between automatic and interactive modes, to support variable selection using statistical ranking measures of correlation to calculate the quality. The interface allows to decompose and visualize the contributions of different data partitions to the quality measure and restrict the selection algorithm to certain subsets of data instances, as well as to select quality measures from a repository and rank the variables accordingly.

Other solutions that rely on well-known statistical measures have been proposed by [Seo and Shneiderman \(2004\)](#), [Seo \(2005\)](#) and [Seo and Shneiderman2 \(2006\)](#). [Seo and Shneiderman \(2004\)](#) describe an interface called Rank-by-feature framework, in which analysts can select a variable ranking criterion and visualize representations of color-coded ranking results such as scatter plots and histograms. Thus, data exploration is guided by the ranking of the variables, individually (ranking of the histograms) or in pairs (ranking of the scatterplots). They have implemented multiple ranking criteria, such as normality of distribution and number of outliers for histograms; or Pearson correlation coefficient and entropy for scatterplots. This framework has been incorporated into a tool called Hierarchical Clustering Explorer (HCE) ([SEO, 2005](#)), for exploratory data visualization. In a later work, [Seo and Shneiderman2 \(2006\)](#) report results from assessing both framework and tool through case studies with users in biology, bioinformatics, and meteorology.

[Piringer, Berger and Hauser \(2008\)](#) later extended the Rank-by-feature framework to consider local variables obtained from several data visualizations through the linking-and-brushing strategy. With a similar goal of exploring the relationships among variables, [Turkay, Filzmoser and Hauser \(2011\)](#) propose a framework with multiple views to explore data instances and variables. [Turkay et al. \(2012\)](#) introduce the construction of representative factors to analyze structures in high-dimensional data. A factor corresponds to a high-level structure that represents subgroups of highly correlated variables. The approach presents selection methods to investigate subsets of variables and associate representative factors to these subsets. The original variables and factors are combined into an interactive visual analysis cycle, facilitating interpretation of the selection process. At each iteration, the factors are refined or generated to provide a better representation of the relationships among variables as a result of individual interactions.

2.3 Visualization and Analysis of Variables in Connection with Specific Mining Tasks

Several papers in the literature deal with the integration of visual representations in variable selection techniques, for input in specific mining tasks, such as clustering, classification or outlier detection. This section introduces contributions that adopt this strategy.

Some studies **focus on classification problems**, such as those by [Zhang *et al.* \(2016b\)](#), [Yang and Moody \(2000\)](#), [Botelho \(2011\)](#) and [Cruz \(2012\)](#). The visualization system by [Zhang *et al.* \(2016b\)](#) supports users to inspect and direct the variable selection in decision trees. A binary tree of variables is generated applying a hierarchical clustering algorithm, whose leaves represent original variables, and internal nodes represent derived variables. The clustering hierarchy reflects different levels of similarity among variables measured by a correlation coefficient. Variables of different tree levels can be selected, serving as input to the decision tree classifier. Users can evaluate the variable selection by analyzing the quality of the resulting classification. Also focusing on classification problems, [Yang and Moody \(2000\)](#) propose to select variables based on the mutual information measure, associated to the data visualization using the Independent Component Analysis (ICA) as dimensionality reduction method. The approach allows to find a variable space that represents the data set, from the analysis of the independent components.

Two contributions by researchers from the Visualization, Imaging and Computer Graphics laboratory (VICG) at ICMC-USP ([BOTELHO, 2011](#); [CRUZ, 2012](#)) use multidimensional projection techniques and Neighbor-Joining (NJ) similarity trees ([CUADROS *et al.*, 2007](#)) to support the identification of meaningful variable spaces. In these cases, the principle is to build a projection or NJ-tree layout of the variables (not the data instances), which can be done simply by transposing the data matrix.

[Botelho \(2011\)](#) projects the feature vectors describing a collection of images – each feature corresponds to a variable. Once the variables are projected there are two possible scenarios. If the projection reveals well-defined clusters, the user can select samples from the different perceived clusters to define a subset of variables, setting up a selection process. Otherwise, if clusters are not easily perceived, an automatic selection algorithm is employed. The projected variables are clustered with K-means to obtain a summary subset of the variable space. Data instances described by distinct variable subsets can be projected again, and the outcome compared with the projection of the original data instances. Projections can be compared qualitatively, through visual inspection, or quantitatively, by means of the Silhouette Coefficient metric ([ROUSSEUW, 1987](#)). Results have shown that this approach for variable selection allows determining variable subsets capable of generating improved projections in terms of visual quality. Also handling image collections, [Cruz \(2012\)](#) uses Neighbor Joining (NJ) similarity trees to select, evaluate and compare variables spaces. The similarity tree generated from the variables allows the user to select a variable space that maintains or improves the quality of the

visualizations of the image set. She proposes a systematic process that supports a structured, user-driven process of comparative analysis of feature extractors, which can be useful in different contexts. Results were evaluated using quality measures such as Silhouette Coefficient and Neighborhood Preservation (PAULOVICH *et al.*, 2008b), as well as the percentage of correctly labeled instances by Multilayer Perceptron and Naive Bayes classifiers.

Dy and Brodley (2000) and Maniyar and Nabney (2006) consider **tasks of unsupervised learning**, where class information is not available to guide the variable selection. Dy and Brodley (2000) describe a platform called Visual Feature Subset Selection using Expectation-Maximization Clustering (Visual-FSSEM), which incorporates clustering and interactive visualization techniques to guide users in identifying relevant variables based on visual perception and evaluation measures. The analyst must define the potentially useful variables of each class, then use them as basis for an automatic selection algorithm that generates a variable subset. In this scenario, the goal is to identify the variable subset to generate the best clustering model, which requires evaluating the quality of different clusters. The solution uses Linear Discriminant Analysis (LDA) to display visual data representations in the two-dimensional space after applying the Expectation-Maximization (EM) clustering algorithm considering distinct subsets of variables. The Maximum-Likelihood criterion is applied to compare clustering model quality. Finally, users can perform both a progressive and a reverse search with backtracking.

In a similar approach, Maniyar and Nabney (2006) modified a method of variable selection in unsupervised learning, which estimates variable saliency simultaneously with the training of the visualization model in chemoinformatics data. The algorithm derives a generative topographic mapping, which is a probability density model that describes the distribution of the data instances in a multidimensional space in terms of a smaller number of latent or hidden variables. The modified algorithm associates variable selection and visualization, provides a better projection by modeling irrelevant variables with a separate noise model, and also gives variable saliency values which help to assess the significance of each variable. Finally, it results in a better separation of data clusters in the layout, providing coherent (compact) projections. Table 1 summarizes important features in the related work reviewed in Sections 2.2 and 2.3.

2.4 Exploratory Visualization of Multivariate Time Series

Time series are a special case of high-dimensional data. They can be defined as sequences of values associated with a temporal order, so-called “raw” time series, which may consist of thousands or millions of values. There are many techniques for visualizing time series, the simplest and oldest being the line graph. However, it does not scale to many multivariate series, as high overlapping impairs interpretation, as illustrated in Figure 4.

Alternative solutions have been introduced for visualizing long time series. For example, the TimeSearcher visualization (HOCHHEISER; SHNEIDERMAN, 2004), illustrated in Figure

Table 1 – Related work on visualization and analysis of variable spaces in multidimensional data.

	Interactivity	User Expertise	Variable Selection	Exploration	Feature Spaces	Clustering	Classification	Ranking	Hierarchy	Coordinated Views
Visualization and analysis of variable spaces										
Yang <i>et al.</i> (2003a)	X	X	X	X		X			X	
Yang <i>et al.</i> (2003b)	X	X	X	X					X	X
Yang <i>et al.</i> (2007)	X		X	X		X				
Schreck, Fellner and Keim (2008)	X				X	X				
Ng and Martin (2004)				X	X					
Yuan <i>et al.</i> (2013)	X			X	X				X	X
Turkay, Filzmoser and Hauser (2011)	X	X		X	X					
May <i>et al.</i> (2011)	X		X	X	X		X			
Seo and Shneiderman (2004)	X			X				X		
Seo (2005)	X			X				X		
Seo and Shneiderman2 (2006)	X			X				X		
Piringer, Berger and Hauser (2008)	X			X				X		
Turkay <i>et al.</i> (2012)	X			X	X	X				
Visualization and analysis of variables in connection with specific mining tasks										
Zhang and Gruenwald (2006)	X		X	X	X		X		X	
Yang and Moody (2000)	X		X	X	X		X			
Botelho (2011)	X		X	X	X	X				
Cruz (2012)	X		X	X	X		X			
Dy and Brodley (2000)	X		X	X		X				
Maniyar and Nabney (2006)	X		X	X		X

Figure 4 – Visualization of 1,430 stock prices along 52 weeks (HOCHHEISER; SHNEIDERMAN, 2004).

5, uses timeboxes to query on entity sets with one or more time-varying variables. Timeboxes are rectangular query regions drawn directly on a 2D line chart depicting the series. The x-axis extension of the timebox indicates the time period, while its y-axis extension restricts the interval of interest. The timebox acts as a filter that only accepts items with values within the defined range. The TimeSearcher tool also contains a graph for each item in the data set, on which queries may be performed by drawing timeboxes.

Figure 5 – TimeSearcher (HOCHHEISER; SHNEIDERMAN, 2004).

Keim, Schneidewind and Sips (2004) consider the representation of both multivariate data flows and historical data. Their CircleView (Figure 6) visualization splits a circle into a number of segments, each representing a variable. Segments are divided into slots, which cover periods of time, and color shows aggregated data over the corresponding interval. Time is mapped linearly along the segments, and the user can adjust the number of time slots interactively.

Several visualizations have been proposed to support **analysis of single or multiple temporal variables**. The Cluster and Calendar based Visualization of Time Series Data (WIJK; SELOW, 1999), shown in Figure 7, presents the data patterns in a calendar and a 3D chart along the time axis. Data is first organized according to a single temporal scale, e.g., daily, weekly, monthly, or other, and then a clustering algorithm is applied to group similar time series. Finally, the patterns identified after clustering are displayed. This technique is limited to data that exhibits temporal regularity and requires prior knowledge of the patterns to be sought. VizTree (LIN *et*

Figure 6 – CircleView (KEIM; SCHNEIDEWIND; SIPS, 2004).

al., 2004) aims to support pattern discovery in data without prior knowledge. The visualization, exhibited in Figure 8, presents a global summary that may reveal hidden structures, of both frequent patterns, which can be detected by thick branches and abnormal patterns, indicated by thin tree branches.

Figure 7 – Cluster and Calendar based Visualization of Time Series Data (WIJK; SELOW, 1999).

The visual approach by Turkay *et al.* (2011) considers two models to enable the exploration and interpretation of structural changes in temporal clusters. The first model, called the temporal clustering view, considers the visual quality of structures of a cluster set over time,

Figure 8 – VizTree (LIN *et al.*, 2004).

which are computed from different time intervals. The structure quality is assessed using the Silhouette Coefficient. The second model, called temporal signatures, represents visual summaries of statistical properties of cluster structures over time. These properties are based on the cohesion and homogeneity of the clusters, which represent the proximity and distribution of the elements. These variations of the statistical properties reveal structural changes in the clusters of elements. These two models are used in an Interactive Visual Analysis (IVA) cycle to analyze temporal clusters, i.e., to observe structural changes such as division, merging, and changes in cluster size. Different connected views are incorporated to interpret and/or evaluate clusters, such as histograms, scatterplots and Parallel Coordinates for regular variables; and graphical functions and animated scatterplots for temporal variables. Figure 9 summarizes the proposed approach, in which a set of temporal clusters is analyzed in terms of structural and quality changes. Poor quality clusters are updated or discarded.

Kumar *et al.* (2005) and Steiger *et al.* (2014) propose solutions for the **visualization of a single time series**. Kumar *et al.* (2005) extract variables from a time series associating them with a relative frequency, and displaying them in a bitmap. In this way, users can visualize similarities and differences in a collection of bitmaps, discovering clusters and anomalies, among other tasks. Figure 10 shows an example of bitmaps relative to four time series.

Steiger *et al.* (2014) present a system to analyze time series generated by networks

Figure 9 – Interactive Visual Analysis (IVA) cycle (TURKAY *et al.*, 2011).

Figure 10 – Bitmap relative to four time series (KUMAR *et al.*, 2005).

of sensors for anomaly detection. The strategy splits records into fixed-length patterns to be compared over time and space, with the support of two linked views. These views allow to cover daily, weekly, or longer periods, with daily patterns being analyzed in a cluster-based view, weekly patterns in a calendar-based view, and patterns relative to longer periods in projection-based views. The connectivity of the sensors can be analyzed through a network topology view, favoring interpretation of the results. This system is shown in Figure 11, and basically consists of two views. The similarity view provides a global overview of all the patterns in the complete network, displaying each pattern as a point in space, where the position is determined by similarity, computed with the Fast Dynamic Time Warping (FDTW) distance. The network view illustrates a diagram of connected nodes associated with on-demand exploration functionalities. The user can thus learn about the distribution of patterns at different levels-of-detail.

Other contributions specifically target the **visualization of time-varying multivariate data**, to guide specialists in particularly challenging tasks that require searching for relevant data patterns in bulky data sets. Several classical multivariate visualization techniques have been employed to convey time-related information, e.g., Parallel Coordinates used in connection

Figure 11 – System by [Steiger et al. \(2014\)](#). The similarity view on the left shows the daily patterns of all sensors, where similar patterns are presented in the same cluster and with the same color. Black lines indicate pattern changes over time, for a specific sensor. The network view on the right provides an overview of the topology, in which each node illustrates changes over time, as well as the patterns in the global context.

with an algorithm to identify important temporal trends in multivariate scientific data sets ([LEE; SHEN, 2009](#)), or combined with histograms to quantify visual properties and reduce visual overload in time-varying volume data ([AKIBAY; MA, 2007](#)).

Figure 12 – Overview of the pipeline of the Correlated Multiples algorithm. (a) Selecting data of interest, analyzing similarity, and finding a similarity-based clustering. (b) The data items are spatially arranged based on similarity, with clusters indicated by color. (c) A proximity graph is computed using Delaunay triangulation to constrain the relative positions of items. (d) The layout is adjusted by the Constrained Multidimensional Scaling (CMDS) algorithm. (e) The final layout is aligned horizontally and vertically ([LIU et al., 2015](#)).

Small multiple visualizations rely on the repetition of the same design structure to summarize and compare variables ([TUFTE, 1991](#)). Correlated multiples ([LIU et al., 2015](#)) is a method that adopts a spatially coherent placement of the multiples views, where their relative distances reflect their dissimilarities. It has been applied to univariate data from three different domains, namely stock market trends, census demographics, and climate modeling, to assess changes over the years and identify common trends and similar years. Figure 12 shows an overview of the pipeline of Correlated multiples. Initially, data of interest are selected for

similarity analysis and finding a similarity-based clustering. A similarity-based layout of the data items is then obtained, where clusters are indicated by color. Then, a proximity graph is computed with the Delaunay triangulation to constrain the relative positions of items. The layout is adjusted by the Constrained Multidimensional Scaling (CMDS) algorithm. Finally, the resulting layout is aligned horizontally and vertically.

TimeSpiral (ZHANG *et al.*, 2016a) is a visualization system combining multiple views to assist users in analyzing and exploring periodic trends and correlations in multivariate time-series data. It supports data aggregation at multiple levels, e.g., modifying a time interval, time span or time granularity. These systems are suitable for simultaneous investigation of a reduced number of variables. The TimeSpiral interface, shown in Figure 13, is organized into three main sections: (1) an overview diagram (region labeled A), (2) a small multiples visualization (B), and (3) a controller composed by the dimension list (C) and the timeline (D). The overview diagram is shown as a radial layout of the multivariate time series data as ring diagrams. The selected period granularity is represented by a ring. Values are represented by curves, where different dimensions are depicted in different colors.

Figure 13 – Interface of the TimeSpiral System. The overview diagram is shown in region A. Region B is the small multiple visualization. The data dimension list is represented in Region C. Finally, region D is the timeline (ZHANG *et al.*, 2016a).

The VIMTEX framework has been designed to assist geologists observing temporal relationships in multivariate data describing concentrations of chemical compounds (DASGUPTA;

Figure 14 – VIMTEX combines multiple visualization components. A time-varying data view of multiple time series allows reordering. A density view shows univariate temporal distributions with the selected pair of axes being highlighted. A matrix view shows bivariate correlations (DASGUPTA; KOSARA; GOSINK, 2015).

(KOSARA; GOSINK, 2015). It uses Parallel Coordinates for a multivariate, time-varying view of the data, combined with a density view to show univariate temporal distribution and a small multiples matrix view, which shows bivariate correlations as time-series. The interface includes three components, as shown in Figure 14: a data view (A) which is the time-varying view of multiple time series and allows a reordering; a density view (B), which shows univariate temporal distribution with the selected axis pair being highlighted; and a matrix view (C) which shows the bivariate correlations.

The interactive system Falcon (STEED *et al.*, 2017) coordinates time-oriented and statistical views for users to explore temporal and statistical patterns in multiple time-varying variables associated with 3D printing processes. In Falcon system, shown in Figure 15, the main window (on the left) consists of interactive visualizations that allow users to explore temporal and statistical patterns in multivariate time series. The variable visualization panel provides an overview and a detailed visualization. After selecting a time range, a statistical summary pane is added to summarize the selected data. Two additional analysis windows are available to provide more detailed views of segments derived from the entire time series of a particular variable.

Our approach bears similarity to the one adopted by Machado *et al.* (2017), who present visual designs based on pixel-oriented layouts, called Player Attribute Heatmap (PAH), to summarize a soccer match using as input players positions over time. The PAH visualization, illustrated in Figure 16, is a matrix display of the player positions during part of the match (often a half-time), conveying how player positions and tactical schemes change along the match. Table 2 summarizes relevant features of the reviewed techniques for visually exploration of time series data sets.

Figure 15 – Falcon system. The main window consists of interactive visualizations that allow users to explore temporal and statistical patterns in multivariate time series. Two additional analysis windows are available to provide more detailed views of segments derived from the entire time series of a particular variable (STEED *et al.*, 2017).

Figure 16 – Player Attribute Heatmap (PAH) visualization of player positions during part of a soccer match, where positions are color-coded conveying team's evolution with respect to attack or defense (MACHADO *et al.*, 2017). Each column represents a player. Dark orange represents defense and dark blue attack.

Table 2 – Features of related work on visualization and analysis of multivariate time series.

	Interactivity	User Expertise	Exploration	Univariate	Multivariate	Clustering	Coordinated Views	Granularity	Dimensionality
Visualization of multivariate time series									
Hochheiser and Shneiderman (2004)			X		X		X		2D
Keim, Schneidewind and Sips (2004)	X		X		X				2D
Wijk and Selow (1999)			X	X	X				2D/3D
Lin <i>et al.</i> (2004)			X	X	X		X		2D
Turkay <i>et al.</i> (2011)	X	X	X	X	X		X		2D
Kumar <i>et al.</i> (2005)				X					2D
Steiger <i>et al.</i> (2014)	X	X	X	X			X		2D
Liu <i>et al.</i> (2015)			X	X		X			2D
Zhang <i>et al.</i> (2016a)	X		X		X		X	X	2D
Dasgupta, Kosara and Gosink (2015)	X		X	X	X	X	X		2D
Steed <i>et al.</i> (2017)	X		X	X	X		X		2D
Machado <i>et al.</i> (2017)	X		X		X		X		2D

2.5 Chapter Notes

In this chapter, we have presented approaches that link visual representations in variable selection, as well as exploring variable spaces in multidimensional data, which help users understand how variables affect observed relationships or patterns. The approaches reviewed support multiple data exploration and mining tasks.

We discussed a few interesting and recent solutions that analyze univariate and multivariate time series, others can be found at: <http://survey.timeviz.net/>. Similarly to the solution proposed by Liu *et al.* (2015), we adopt a small multiples visualization and a consistent placement strategy to display groups of related variables; and in addition, we consider arbitrary time periods of different granularity levels by means of user-defined data aggregations (ZHANG *et al.*, 2016a), which allows visual data explorations at different detail levels.

Although some reviewed work visualize single variables or collections, only the solutions proposed by Turkay, Filzmoser and Hauser (2011), Turkay *et al.* (2011) and Yuan *et al.* (2013) explore the similarity and relationships between both variables and/or data instances. Moreover, none of the previous strategies is directly applicable to the tasks initially identified in connection with the scintillation data because they do not suffice to answer all the ionospheric data analysis questions. Our novel framework *TV-MV Analytics* advances previous efforts by providing multiple and complementary strategies to inspect and compare the historical behavior of time-varying multivariate data, as described in the following chapters.

TIME-VARYING MULTIVARIATE DATA SETS

3.1 Initial Remarks

A time-varying multivariate data set comprises multivariate data items defined over a temporal range, wherein each data instance is a time-stamped multivariate observation, i.e. recording multiple variables. A simple example is illustrated in Table 3, where each row describes a time-stamped data instance and each column describes a variable. The data instances describe year-average measurements of *age*, *height*, *weight*, and *body mass index* (BMI) of a Pakistani boy, recorded from 1997 to 2002. Data has been extracted from a pediatric study of school children of different socio-economic and cultural backgrounds (AZIZ *et al.*, 2012).

Table 3 – A time-varying multivariate data set, where each row describes a time-stamped data instance and each column describes a particular variable. Data instances describe year-average physiological measures of a Pakistani boy from 1997 to 2002 (AZIZ *et al.*, 2012).

Age	Height	Weight	BMI	Year
1	49.2	5.20	23.0	1997
2	77.5	6.1	25.7	1998
3	89.8	12.0	21.2	1999
4	94.1	13.1	16.5	2000
5	104.0	16.2	19.8	2001
6	109.8	17.3	16.1	2002
7	116.9	18.8	15.9	2003
8	121.9	21.8	15.9	2004
9	128.0	23.6	16.0	2005
10	139.1	26.2	16.0	2006
11	141.1	30.0	15.6	2007
12	147.2	33.0	16.6	2008
13	151.9	38.1	17.5	2009
14	157.8	40.9	17.7	2010
15	158.5	45.6	18.3	2011
16	162.7	45.7	18.3	2012

Analyzing time-varying multivariate data may require evaluating clusters, correlations, and other features, where considering the temporal aspect is an essential requirement. For instance, in Table 3, it might be of interest (i) to find the relationships, if any, between the behaviors of the multiple variables, (ii) to study the behavior of the variables over the years, or (iii) to study how the behavior of multiple variables changes in different individuals.

This chapter describes the time-varying multivariate data sets considered in our experimental results, originated from four different domains: ionospheric scintillation, which provided the major motivation for this research (Section 3.2), and three additional data sets, namely on crime statistics of the state of São Paulo, Brazil (Section 3.3.1), on air quality measurements at an Italian city (Section 3.3.2), and on US stock market (Section 3.3.3).

3.2 Ionospheric Scintillation Data

This doctoral project involved a collaboration with researchers from UNESP at Presidente Prudente who study the effects of ionospheric scintillation. The ionosphere is the upper layer of the atmosphere, and because it is permanently ionized, it acts as a propagation medium, enabling satellite communication by means of high-frequency radio waves around the world. Services such as the Global Navigation Satellite System (GNSS) allow determining geospatial positioning using satellite signals.

The ionospheric scintillation is an atmospheric phenomenon that results from amplitude attenuation and phase shifts in GNSS radio signals as they propagate through regions of the ionosphere with irregular electron densities (YEH; LIU, 1982). It affects measurements obtained by GNSS systems, such as the Global Positioning System (GPS, United States), presenting a discontinuity in the phase tracking loop (ALFONSI *et al.*, 2011), as well as a degradation in the accuracy of the measurements from a given satellite (KIEFT; AQUINO; DODSON, 2014). The occurrence of ionospheric scintillation impairs good communication in the reception of satellite signals, whereby the received information presents spurious and/or absent data.

Effects vary by signal frequency, geographic location, time, season, magnetic activity, and solar cycles. Figure 17 shows the frequency of occurrence of ionospheric scintillation according to geographic location. It can be observed that frequency is higher in the central region, but even there, there is a small band with lower frequency where the magnetic equator is located. In the equatorial ionosphere, there is a combination of the electric and magnetic fields, causing the release of electrons vertically, diffused to the North and to the South. Thus, ionization is reduced in the equator (VANI *et al.*, 2013).

There is enormous interest in analyzing the effects of scintillation and understanding its causes in order to gather elements to deal with its effects. In Brazil, GNSS receivers have been installed in different stations in the context of the European Union-funded CIGALA project (ALFONSI *et al.*, 2011) in order to collect data relative to this phenomenon. These receivers

Figure 17 – Frequency of occurrence of ionospheric scintillation according to geographic location (KINTNER; HUMPHREYS; HINKS, 2009).

record 1,440 measurements daily per satellite tracked, which are stored in a database that contains an expressive volume of measures and provides an important source of information.

Figure 18 shows the geographical location of the stations available since July 2013 and probable future stations. Two stations located at Presidente Prudente (PRU1 and PRU2) transmit data to a repository at FCT/UNESP, which provided the data analyzed in this work.

Figure 18 – Stations of the CIGALA/CALIBRA Network (VANI *et al.*, 2013).

3.2.1 Indicators and Measurements of Ionospheric Scintillation

Global Navigation Satellite Systems (GNSS) consist of satellite constellations that allow determining the coordinates of a receiver's antenna. At present, there are four major GNSS

constellations: the North American GPS (Global Positioning System), the Russian GLONASS (GLObal'naya NAvigatsionnaya Sputnikkovaya System), the European GALILEO and the Chinese Beidou/Compass. GNSS also includes complementary systems, such as those based on geostationary satellites and ground stations. The stations collect measurements every minute, each record described by 70 variables computed/calculated (see Appendix A) and data instances are generated every minute sampled at high rates (50 Hz), for each satellite of the aforementioned constellations tracked by the receivers.

The variables can be interpreted as multivariate time series of numerical values describing multiple scintillation indices (43 variables), measures of the S4 index (12 variables), measures of the SigmaPhi index (6 variables), time stamps (7 variables), and indicators of a satellite's spatial orientation (2 variables).

- **Identifiers:** *sbfblock*, *sbf2ismr* and *svid* were not considered for data analysis purposes (Table 4). *sbfblock* informs the state of the receiver, *sbf2ismr* its version number and *svid* identifies the satellite.

Table 4 – Identifiers considered in the ionospheric scintillation data (Septentrio Satellite Navigation, 2013).

#	Variable	Description	Category
1	<i>svid</i>	Satellite id	identifier
2	<i>sbfblock</i>	Value of RxState of the Receiver Status	identifier
3	<i>sbf2ismr</i>	Sbf2ismr version number	identifier

- **Diverse scintillation indices:** describe the standard deviation of phase computed over different time intervals; average and standard deviation of code/carrier divergence; spectral strength and spectral slope of detrended phase; averaged signal-to-noise ratio; and absolute and differential total electron content (Table 5).

Specifically, the variables *avg_cn0_** report the C/N0 average of the different last minute signals in dB-Hz. The variables *avgccd_** and *sigmaccd_** report the mean value and the standard deviation of the difference between code and phase measured from the last minute for each of the signals.

The variable *f2nd_tec_cn0* reports the average C/N0 (in dB-Hz) in the second frequency, used to calculate the TEC, which gives the total electron content in a satellite-receiver trajectory. DTEC captures variations of TEC at intervals of 15, 30, 45 and 60 seconds. Both TEC and DTEC are given in TECU units: a TECU unit is equivalent to 10^{16} electrons/ m^2 (VANI, 2014).

Variables *phi_** report the standard deviation of the normalized phase by averaging the values at intervals of 01, 03, 10, 30 and 60 seconds, respectively, for each of the signals L1, L2 and L3. There are also measures related to the spectral slope (*p_**) and strength

Table 5 – Diverse scintillation indices that describe the behavior of ionospheric scintillation data ([Septentrio Satellite Navigation, 2013](#)).

#	Variable	Description	Category
1	<i>avg_cn0_l1</i>	Average L1 C/N0 over last minute	index
2	<i>avg_cn0_l2</i>	Average L2 C/N0 over the last minute	index
3	<i>avg_cn0_l5_e5b</i>	Average L3 C/N0 over the last minute	index
4	<i>avgccd_l1</i>	Average of code/carrier divergence	index
5	<i>avgccd_l2c_e5a</i>	Average of code/carrier divergence on L2	index
6	<i>avgccd_l5_e5b</i>	Average of code/carrier divergence on L3	index
7	<i>sigmaccd_l1</i>	Standard deviation of code/carrier divergence	index
8	<i>sigmaccd_l2c_e5a</i>	Standard deviation of code/carrier divergence on L2	index
9	<i>sigmaccd_l5_e5b</i>	Standard deviation of code/carrier divergence on L3	index
10	<i>f2nd_tec_cn0</i>	Averaged C/N0 of 2nd frequency	index
11	<i>tec</i>	TEC at TOW	index
12	<i>tec_15</i>	TEC at TOW - 15 seconds	index
13	<i>tec_30</i>	TEC at TOW - 30 seconds	index
14	<i>tec_45</i>	TEC at TOW - 45 seconds	index
15	<i>dtec_15tow</i>	DTEC from TOW - 15s to TOW	index
16	<i>dtec_3015</i>	DTEC from TOW - 30s to TOW - 15s	index
17	<i>dtec_4530</i>	DTEC from TOW - 45s to TOW - 30s	index
18	<i>dtec_6045</i>	DTEC from TOW - 60s to TOW - 45s	index
19	<i>phi01l1</i>	Phi01 on L1 - 1-second phase sigma (radians)	index
20	<i>phi03l1</i>	Phi03 on L1 - 3-second phase sigma	index
21	<i>phi10l1</i>	Phi10 on L1 - 10-second phase sigma	index
22	<i>phi30l1</i>	Phi30 on L1 - 30-second phase sigma	index
23	<i>phi60l1</i>	Phi60 on L1 - 60-second phase sigma	index
24	<i>phi01_l2c_e5a</i>	Phi01 on L2-1-second phase sigma	index
25	<i>phi03_l2c_e5a</i>	Phi03 on L2-3-second phase sigma	index
26	<i>phi10_l2c_e5a</i>	Phi10 on L2 - 10-second phase sigma	index
27	<i>phi30_l2c_e5a</i>	Phi30 on L2 - 30-second phase sigma	index
28	<i>phi60_l2c_e5a</i>	Phi60 on L2 - 60-second phase sigma	index
29	<i>phi01_l5_e5b</i>	Phi01 on L3 - 1-second phase sigma	index
30	<i>phi03_l5_e5b</i>	Phi03 on L3 - 3-second phase sigma	index
31	<i>phi10_l5_e5b</i>	Phi10 on L3 - 10-second phase sigma	index
32	<i>phi30_l5_e5b</i>	Phi30 on L3 - 30-second phase sigma	index
33	<i>phi60_l5_e5b</i>	Phi60 on L3 - 60-second phase sigma	index
34	<i>p_l1</i>	P (spectral slope of detrended phase) on L1	index
35	<i>p_l2c_e5a</i>	P (phase spectral slope) on L2	index
36	<i>p_l5_e5b</i>	P (spectral slope of detrended phase) on L3	index
37	<i>t_l1</i>	T (phase power spectral density) on L1	index
38	<i>t_l2c_e5a</i>	T (phase power spectral density) on L2	index
39	<i>t_e5b</i>	T (phase power spectral density) on L3	index
40	<i>vphi60_l1</i>	Verticalization of Phi60 on L1	index
41	<i>vphi60_l2c_e5a</i>	Verticalization of Phi60 on L2	index
42	<i>vphi60_l5_e5b</i>	Verticalization of Phi60 on L3	index
43	<i>vtec</i>	Verticalization of TEC	index

(t_*). Variables v_* refers to the verticalization of other variables, which means the relative values on a perpendicular to the ground standing route.

- **S4 indices:** measure the scintillation amplitude in different bands, as the standard deviation of the detrended signal intensity (total S4), considering noise effects (multipath and others), the thermal noise correction of the scintillation amplitude (S4 correction), the corrected indices (corrected S4), and the verticalization of the total S4 index (Table 6).

Table 6 – S4 indices related to the behavior of ionospheric scintillation of amplitude ([Septentrio Satellite Navigation, 2013](#)).

#	Variable	Description	Category
1	$s4$	Total S4 on L1	s4_index
2	$s4_{l2}$	Total S4 on L2	s4_index
3	$s4_{l5_e5b}$	Total S4 on L3 (dimensionless)	s4_index
4	$corr_s4$	Correction to total S4 on L1	s4_index
5	$corr_s4_{l2}$	Correction to total S4 on L2	s4_index
6	$corr_s4_{l5_e5b}$	Correction to total S4 on L3	s4_index
7	$s4_corrected$	Total S4 on L1 corrected	s4_index
8	$s4_{l2_corrected}$	Total S4 on L2 corrected	s4_index
9	$s4_{l5_e5b_corrected}$	Total S4 on L3 corrected	s4_index
10	$vs4_{l1}$	Verticalization of S4 on L1	s4_index
11	$vs4_{l2}$	Verticalization of S4 on L2	s4_index
12	$vs4_{l5}$	Verticalization of S4 on L5	s4_index

Variables $s4_*$ are the standard deviation of observations¹, free of trend signal intensity, sampled at high rates (e.g. 50 Hz) at a range of 60 seconds on the different signals. For obtaining these indices the signal power is normalized with respect to the received mean power ([VANI *et al.*, 2013](#)).

The S4 index of scintillation amplitude is often employed to measure the intensity of ionospheric scintillation. It is computed as the standard deviation of the signal intensity normalized by its mean ([DIERENDONCK; KLOBUCHAR; HUA, 1993](#)). Signal intensity must be measured at high rates for the index to detect rapid fluctuations.

The S4 total indices are defined by Equation 3.1:

$$S4_{total} = \sqrt{\frac{\langle SI^2 \rangle - \langle SI \rangle^2}{\langle SI \rangle^2}}. \quad (3.1)$$

where SI is the signal intensity sampled at 50 Hz and the operator $\langle \rangle$ indicates the average over a one-minute interval.

¹ Observations are basic measurements of a common GNSS receiver.

Variables $corr_s4_*$ report the value of $S4$ from ambient noise for each signal obtained by estimating the mean signal-to-noise density over the one-minute interval in different bands. The correction of the $S4$ index uses these values, as described in Equation 3.2.

$$X = S4_{total}^2 - S4_{correction}^2$$

$$S4_{corrected} = \begin{cases} \sqrt{X} & \text{se } X > 0 \\ 0 & \text{se } X \leq 0 \end{cases} \quad (3.2)$$

Currently, there is no consistency concerning the categorization of the severity of scintillation, as measured by the $S4$ index. Conker *et al.* (2003) state that an $S4$ index value of 0 indicates absence of scintillation while a value 1 indicates severe scintillation. Tiwari *et al.* (2011) proposes a categorization into three levels: weak, moderate and strong, which is one of the most used by the specialists. Hegarty *et al.* (2001) state that the upper limit for the $S4$ index is 1.414 and propose a categorization in four levels, whereas ITUR (2013) defines three levels. Table 7 presents the categories of $S4$ according to different authors.

Table 7 – Categorization of the $S4$ index according to different authors.

S4 Categorization		
<i>Tiwari et al.</i> (2011)		
Category	Range	Level
1	$S4 \geq 1$	high
2	$0,5 \leq S4 < 1$	moderate
3	$S4 < 0,5$	low
<i>Hegarty et al.</i> (2001)		
Category	Range	Level
1	$S4 \geq 0,9$	high
2	$0,6 \leq S4 < 0,9$	moderate
3	$0,4 \leq S4 < 0,6$	low
4	$0,1 \leq S4 < 0,4$	very low
<i>ITUR</i> (2013)		
Category	Range	Level
1	$S4 \geq 0,6$	high
2	$0,3 < S4 < 0,6$	moderate
3	$S4 \leq 0,3$	low

In general, the receivers can provide values higher than the theoretical limits. For example, the multipath effect may increase the $S4$ index, falsely indicating the incidence of ionospheric scintillation. This problem occurs mainly at the beginning, or at the end, of the tracking of a satellite, characterized by low elevation angles, consequently more susceptible to the multipath effect. Techniques for removing or reducing such effects can be employed, such as using the standard deviation of the code/phase divergence. Also variables $vs4_*$ refers to the verticalization of $s4$.

- **Sigma-phi indices:** measure phase scintillation in different bands as the standard deviation of the detrended carrier phase from received GPS signals (Table 8).

Table 8 – Sigma-phi indices related to the behavior of ionospheric scintillation of phase ([Septentrio Satellite Navigation, 2013](#)).

#	Variable	Description	Category
1	<i>si_l1</i>	SI Index on L1	si_index
2	<i>si_l2c_e5a</i>	SI Index on L2	si_index
3	<i>si_l5_e5b</i>	SI Index on L3	si_index
4	<i>numerator_si_l1</i>	SI Index on L1 - numerator only	si_index
5	<i>numerator_si_l2c_e5a</i>	SI Index on L2, numerator only	si_index
6	<i>numerator_si_l5_e5b</i>	SI Index on L3, numerator only	si_index

The variables *si_** indicate the sigma-phi indices in each signal for phase scintillation. The sigma-phi indices consist of the standard deviation of trend-free observations of the signal phase, which are collected at high frequencies (for example, 50 Hz) in a range of one minute, and is defined in Equation 3.3.

$$\sigma - \phi = \sqrt{\langle SI \rangle^2 - \langle SI^2 \rangle}, \quad (3.3)$$

with SI defined as:

$$SI = \frac{10 \log P_{max} - 10 \log P_{min}}{10 \log P_{max} + 10 \log P_{min}}, \quad (3.4)$$

in which P_{max} refers to the largest power amplitude in the path of 3,000 samples over the last minute, and P_{min} is the smallest power amplitude.

The range of sigma-phi index values may go from 0.05, for weak scintillations, up to 1.00 on strong scintillations ([SKONE et al., 2005](#)). Table 9 presents categorizations of the sigma-phi index proposed by [Tiwari et al. \(2011\)](#) and [Hegarty et al. \(2001\)](#).

Table 9 – Categorization of sigma-phi index according to different authors.

Sigma-phi Categorization		
Tiwari et al. (2011)		
Categorization	Range	Level
1	$\sigma - \phi \geq 0,8$	high
2	$0,4 \leq \sigma - \phi < 0,8$	moderate
3	$\sigma - \phi < 0,4$	low
Hegarty et al. (2001)		
Categorization	Range	Level
1	$\sigma \phi \geq 0,6$	high
2	$0,3 \leq \sigma \phi < 0,6$	moderate
3	$0,2 \leq \sigma \phi < 0,3$	low
4	$0,05 \leq \sigma \phi < 0,2$	very low

The variables *numerator_si_** report the modification of *SI* index, considering just the numerator.

- **Time stamps:** The time stamps consist of the lock time (synchronization time) between satellites and receivers in different bands and GPS time (Table 10).

Table 10 – Variables related to the synchronization time and GPS time of ionospheric scintillation (Septentrio Satellite Navigation, 2013).

#	Variable	Description	Category
1	<i>wn</i>	GPS week number	time
2	<i>tow</i>	GPS time of week	time
3	<i>dow</i>	Day of week	time
4	<i>f2nd_tec_locktime</i>	Lock time on 2nd frequency	time
5	<i>l1_locktime</i>	L1 lock time (seconds)	time
6	<i>l2_e5a_locktime</i>	L2 lock time	time
7	<i>l5_e5b_locktime</i>	L3 lock time	time

The variables *wn*, *tow* and *dow* report the week number, the time of week and the day of week in the GPS scale.

Variables **_locktime* indicate the lock time of the different signals. *F2nd_tec_locktime* report the lock time (in seconds) on the second frequency, used to calculate the TEC.

- **Spatial orientation:** The indicators of a satellite's spatial orientation refer to satellite azimuth and elevation angles relative to the receiver (Table 11).

Table 11 – Variables related to the spatial orientation of ionospheric scintillation (Septentrio Satellite Navigation, 2013).

#	Variable	Description	Category
1	<i>azim</i>	Azimuth	direction
2	<i>elev</i>	Elevation	direction

The variable *azim* represents the azimuth, a measure of angular aperture defined in degrees, ranging from 0° to 360°; *Elev* is the satellite elevation angle.

3.2.2 Previous Work on Visualizing Ionospheric Scintillation Data

The ISMR Query Tool (VANI *et al.*, 2013; VANI, 2014) is an integrated software platform specifically developed for the Brazilian scintillation database. Most of its visualizations are univariate: point-based visualizations are employed to support tasks such as inspecting a single variable or comparing two variables. Figure 19 illustrates different examples: a one-week visualization of GPS data from the SJCU station (Figure 19.a) and comparison of measurements from two stations (Figure 19.b).

It also offers a grid-based visualization, exemplified in Figure 20, which shows the *s4* index values relative to all GPS satellites from the PRU1 station over a period of three hours.

Figure 19 – Examples of point-based visualizations in the ISMR Query Tool (VANI, 2014) (UNESP)

Each line represents a particular satellite, and the columns represent the time sequence; the color of the cells maps the values (black indicates missing values). In this visualization the information does not overlap, allowing to identify the availability of the satellites over time in a station.

Figure 20 – Grid-based visualization: The s4 index relative to all GPS satellites from station PRU1 (VANI, 2014).

A map visualization supports the evaluation of the indices through pierce points projected on a map of the Brazilian territory, using different colors for mapping the variables. The pierce points are positioned on the map and currently display: mean, standard deviation, minimum, maximum, and the ratio of the number of pierce points with a value greater or less than a user-defined threshold. Figure 21 illustrates this view, showing the number of pierce points with S4 greater than 0.5 in the second half of 2011, for station PRU1.

There is also a horizon charts visualization, called Cubism, which allows inspecting up to tens of variables simultaneously over a daily period. The x-axis consists of 1,440 pixels, in which each pixel represents one minute. Figure 22 shows an example displaying 6 variables simultaneously.

Finally, a visualization of series of values uses the approach known as Symbolic Aggregate Approximation (SAX) (YI; FALOUTSOS, 2000), which reduces the dimensionality of a

Figure 21 – Pierce points visualization (VANI, 2014).

Figure 22 – Visualizing the behavior of 6 variables on the horizon charts (VANI, 2014).

time series and generates a symbolic representation. The first step is to reduce the time series using the Piecewise Aggregate Approximation (PAA) technique (LIN *et al.*, 2003), which splits a time series of dimensionality m into fixed-length segments. Then, each segment is represented by its average value. Symbols are then assigned to each segment resulting from the PAA through the use of breakpoints. Finally, the time series is converted into a symbolic representation with SAX, shown in Figure 23, in which the results of the SAX technique are visualized for the 30-day values of the s4 index using 15-minute windows of all GPS satellites tracked by station PRU1. Each line shows at the beginning the satellite identifier and at the end the time windows considered.

The so-called S4 index of scintillation amplitude is often employed to measure the intensity of ionospheric scintillation, as mentioned in Chapter 3. Several authors applied data mining techniques to collected scintillation data, e.g., Rezende *et al.* (2010) devise a method that combines a bagging method, which uses bootstrap to randomly generate several samples

Figure 23 – Visualization using the SAX technique (VANI, 2014).

from an original sample, with decision trees to predict the S4 index with one or more hours of antecedence. Lima *et al.* (2014) present correlation analysis of occurrences of ionospheric scintillation registered in two stations at different locations in Brazil. They used a Classification and Regression Decision Tree (CART) from the S4 index. Another contribution (LIMA *et al.*, 2015) relies on neural networks to predict scintillation at two levels: strong or not strong (low/moderate). Analyses are typically supported by simple univariate time series visualizations.

Figure 24 – Daily variation of S4 index in a West African equatorial region from January 2008 to December 2008 with an outage of data in August (ACKAH; OBROU; GROVES, 2011).

Ackah, Obrou and Groves (2011) investigate records of the S4 and vertical TEC (VTEC) indices in a West African equatorial region. Their method visually represents the time series as grids, where a column represents a daily hour, a row represents a particular day, and grid cell color maps an index (S4 or VTEC) at the corresponding day/hour. The most highlighted

feature of this visualization is that the variable behaviors and its patterns can be perceived clearly. Therefore, we have similarly adopted grid representations to convey variable values measured over a time range, as introduced in Chapter 5. Figure 24 shows the proposed grid with the daily variation of S4 index in a West African equatorial region from January 2008 to December 2008 with an outage of data in August.

3.3 Other Data Sets

The applicability of the framework developed has also been illustrated on three additional data sets from other domains, described next.

3.3.1 Crime Statistics for the State of São Paulo, Brazil

Crime statistics provides information on public safety and allow planning police actions and investments. In the state of São Paulo, the Secretariat of Public Security is responsible for data collection and analysis to support crime prevention and repression. A public crime statistics data set is available² describing several indicators of criminal occurrences in 12 areas indicated in the map in Figure 25, recorded monthly from January 1, 2002 to February 1, 2017.

Figure 25 – Map of the regions for which crime statistics are recorded in State of São Paulo, Brazil.

The data instances comprise 13 variables recording the numbers of monthly occurrences related to³: (1) possession of narcotics; (2) traffic of narcotics; (3) seizure of narcotics; (4) criminal possession of weapons; (5) seized firearms; (6) official reports of flagrant crimes; (7)

² Secretariat of Public Security, Government of the State of São Paulo, Brazil, Police Productivity <<http://www.ssp.sp.gov.br/Estatistica/Pesquisa.aspx>>, (December 27, 2017)

³ We have translated the names of offenses following the US common legal definitions provided at <<https://uslegal.com>> (free translation).

detentions of flagrant offenders; (8) detentions of offenders by warrant; (9) arrests of flagrant offenders; (10) arrests of offenders by warrant; (11) imprisonments; (12) recovered vehicles; and (13) total inquiries. We added a variable (14) informing the homicide rates over the same period, also reported by the state Secretariat of Public Security⁴. All indicators are listed in Table 12

Table 12 – Crime indicators

#	Indicator
1	possession of narcotics
2	traffic of narcotics
3	seizure of narcotics
4	criminal possession of weapons
5	seized firearms
6	official reports of flagrant crimes
7	detentions of flagrant offenders
8	detentions of offenders by warrant
9	arrests of flagrant offenders
10	arrests of offenders by warrant
11	imprisonments
12	recovered vehicles
13	total inquiries
14	homicide rates

Some indicators, such as a number of seized firearms, occurrences of narcotics traffic and number of imprisonments are considered as indicators of police activity. Hence, an increase in their occurrence is interpreted as a positive indicator of stronger police action.

This data on crime statistics has been investigated by [Arvate and Souza \(2011\)](#), who estimated the impact of fire-armed police on crime in Brazil after a law was approved to regulate firearm usage by municipal police forces. Authors constructed indicators for each state region for years 2002, 2004, 2006, 2009 and 2012, using variables: total number of inquiries, individuals arrested in flagrant, drug seizure, recovered vehicles, prisons issued and homicide. They created regression functions using these indicators as dependent variables, and concluded that armed municipal police forces contributed to a reduction in crime indicators after 2003; they also attribute to this factor a significant increment observed in the number of individuals arrested in flagrant.

3.3.2 Air Quality Data at an Italian City

The third data set⁵ describes measurements from urban environment pollution monitoring. Measurements were obtained from two sensing devices, deployed at an Italian city with heavy car traffic, from March 10, 2004 to April 04, 2005. Six pollutant gases are tracked (some of them

⁴ Secretariat of Public Security, Violent Crime Rate <<http://www.ssp.sp.gov.br/Estatistica/Pesquisa.aspx>>, (December 27, 2017)

⁵ UCI, Air Quality Data Set, <<https://archive.ics.uci.edu/ml/datasets/Air+quality>>, (December 27, 2017)

by both devices): carbon monoxide CO , non metanic hydrocarbons $NMHC$, nitrogen oxide NO_x , nitrogen dioxide NO_2 , trioxygen/ozone O_3 , and benzene C_6H_6 .

One of the sensing devices is referred to as the MOX sensor, as it is equipped with five Metal Oxides (MOX) sensors (VITO *et al.*, 2008), namely: (i) a tin oxide sensor to estimate carbon monoxide (MOX_CO), (ii) a titania sensor to estimate non metanic hydrocarbons (MOX_NMHC), (iii) a tungsten oxide sensor to estimate nitrogen oxide (MOX_NO_x), (iv) a tungsten oxide sensor to estimate nitrogen dioxide (MOX_NO_2) and (v) a indium oxide to estimate trioxygen/ozone (MOX_O_3); their concentrations are recorded hourly. The other sensing device is a conventional air pollution monitoring station equipped with spectrometer analyzers to provide reference data. The fixed station provided concentration estimations of five atmospheric pollutants: carbon monoxide (ST_CO), non metanic hydrocarbons (ST_NMHC), benzene ($ST_C_6H_6$), nitrogen oxide (ST_NO_x) and nitrogen dioxide (ST_NO_2). Two commercial temperature and humidity sensors were co-located with pollution monitoring sensors.

The data instances are thus described by 13 variables measured (LIVINGSTON; DECKER; AI, 2013) by these devices (hourly averages), as listed in Table 13.

Table 13 – Variables of UCI air quality data set (VITO *et al.*, 2008).

	Variable	Description	Device	Unit
1	ST_CO	carbon monoxide	spectrometer analyzer	mg/m^3
2	MOX_CO	carbon monoxide	MOX sensor - tin oxide	mg/m^3
3	ST_NMHC	non metanic hydrocarbons	spectrometer analyzer	$\mu g/m^3$
4	$ST_C_6H_6$	benzene	spectrometer analyzer	$\mu g/m^3$
5	MOX_NMHC	non metanic hydrocarbons	MOX sensor - titania	ppb
6	ST_NO_x	nitrogen oxide	spectrometer analyzer	ppb
7	MOX_NO_x	nitrogen oxide	MOX sensor - tungsten oxide	$\mu g/m^3$
8	ST_NO_2	nitrogen dioxide	spectrometer analyzer	$\mu g/m^3$
9	MOX_NO_2	nitrogen dioxide	MOX sensor - tungsten oxide	ppm
10	MOX_O_3	trioxygen/ozone	MOX sensor - indium oxide	$\mu g/m^3$
11	T	temperature	commercial sensor	$^{\circ}C$
12	RH	relative humidity	commercial sensor	%
13	AH	absolute humidity	commercial sensor	%

This data set has been investigated by Vito *et al.* (2008) and Vito *et al.* (2009). Vito *et al.* (2008) consider the problem of estimating Carbon Monoxide (CO), Nitrogen Oxide (NO_x), Nitrogen dioxide (NO_2) and Benzene (C_6H_6) concentrations, from the MOX sensor and/or the conventional station responses. They reported benzene-related findings, which has no specific sensor in the MOX device, and found significant correlations among some variables over the period. They employed back-propagation neural networks to predict benzene (C_6H_6) concentrations, with low estimation errors, and found that the NMHC pollutant has a notable benzene-related component.

Vito *et al.* (2009) study the influence of variable selection and correlation analysis in estimating CO , NO_x and NO_2 concentrations also using back-propagation neural networks. The variable selection task considers all possible feature set combinations for estimating concentrations of NO_2 , CO and NO_x , evaluating alternative subsets of variables with separate regressors and a fixed training set length of 10 days. Interestingly, the best performance values in estimating NO_2 have been obtained using all the sensors responses; e.g. using NO_x and NO_2 sensors responses, CO and Non-Methane Hydrocarbons ($NMHC$) sensors, or NO_x , NO_2 and CO . They confirm the influence of temperature and humidity when NO_x and NO_2 sensors are concerned. Besides, results on CO estimation revealed that $NMHC$ can follow the concentration of CO significantly, and that results deteriorate when NO_2 and NO_x are added.

3.3.3 Stock Market Data

The stock market data set has been extracted from a platform for financial and economic data, called Quandl⁶. We considered data on equities from the segment of technological companies: the 12 leaders in the sector and all 438 technological companies in US, from January 2, 1962 to November 30, 2016, which are described by seven variables recorded daily on stock behavior, namely: (1) *open*, (2) *close*, (3) *high*, (4) *low*, (5) *volume*, (6) *ex_dividend* and (7) *split ratio*.

Variable *open* refers to the stock's opening price at a particular day. *Close* is the stock's closing price. *High* and *low* are the stock's highest and lowest prices, respectively. *Volume* is the number of shares or contracts traded during a day. *Ex_dividend* informs profits distributed to shareholders. The *ex-dividend* is zero if the upcoming dividend is given to the buyer. *Split ratio* is the ratio of shares compared to what was available prior to a stock split (when existing shares are split into multiple shares). Usually, investors use the *split ratio* to determine how much a stock was split, where a higher split ratio indicates a lower share price.

The original data set contains information on 3,000 companies, which we classified into 12 categories, following the New York Stock Exchange⁷ and the Nasdaq Stock Market⁸: (1) capital good; (2) energy; (3) technology; (4) public utilities; (5) basic industries; (6) finance; (7) consumer-non-durable; (8) consumer-durable; (9) transportation; (10) miscellaneous (conglomerated or multi-sector companies); (11) health care; and (12) consumer services.

Several authors proposed visualizations to extract insights and assist in decision making (KO *et al.*, 2016) in stock market data analysis. Popular visualizations are line graphs (RUDOLPH; SAVIKHIN; EBERT, 2009) and candlestick charts (EDWARDS; BASSETTI; MAGEE, 2007), which are usually employed to represent regressions. Aggregation methods

⁶ Quandl, Wiki EOD Stock Prices, <<https://www.quandl.com/data/WIKI/documentation/bulk-download>>, (December 27, 2017)

⁷ New York Stock Exchange, <<https://www.nyse.com/>>, (December 27, 2017)

⁸ Nasdaq Stock Market, <<http://www.nasdaq.com/>>, (December 27, 2017)

are common in this context, to reduce overplotting and occlusion, although they may induce information loss (SORENSEN; BRATH, 2013) at the expense of providing a data overview. A common mining task to identify patterns is clustering (KO *et al.*, 2016), and several algorithms have been applied to stock market data, such as k -means (ZIEGLER *et al.*, 2010) and a variant called k -core (LEI; ZHANG, 2010). Figure 26 exemplifies a visualization proposed by Ziegler *et al.* (2010) in which segments of 550 stocks, from May 1994 to June 2010, were clustered. The stocks from software and banking companies are represented in yellow and in brown, respectively. The identified patterns, shown in each column header, summarize the behavior of stocks from both sectors. It is possible to notice that the patterns associated with companies in the technology sector are considerably distinct, in general, from those in the banking sector.

Figure 26 – Clusters of segments of 550 stocks from May 1994 to June 2010, where companies in the software sector are shown in yellow and companies in the banking sector are shown in brown (ZIEGLER *et al.*, 2010).

3.4 Chapter Notes

In this chapter we have described the nature of the time-varying multivariate data sets employed to validate the *TV-MV Analytics* framework illustrating its applicability on additional exploratory tasks. The data sets originated from four different domains: ionospheric scintillation (the major motivation of this research), crime statistics of the state of São Paulo-Brazil, air quality of an Italian city, and US stock market. Table 14 summarizes relevant characteristics of these data sets.

Table 14 – Data sets analyzed.

Domain	# Var.	Rate	# Instances	Initial date	Final date	Size
Ionospheric Scintillation	70	minute	10,769,507	2014-01-01	2015-12-31	5GB
Crime Statistics	14	month	2,366	2002-01-01	2017-02-01	221KB
Air Quality	13	hour	9,357	2004-03-10	2005-04-04	830KB
Stock Market	7	day	1,773,230	1995-01-01	2016-11-30	2GB

THE VISUAL ANALYTICS FRAMEWORK

“The purpose of computing is insight, not numbers nor pictures.”

(R. Hamming and L. N. Trefethen)

4.1 Initial Remarks

The challenges inherent in exploring time-varying multivariate data are critical in many application domains and require specific approaches and tailored solutions. We worked with collaborators, experts in the ionospheric scintillation phenomenon in order to identify the requirements and the most important issues for ionospheric data analysis. The specifications suggest that analysis has to consider (i) the temporal behavior of variables individually and in groups, as well as (ii) the relationships of different subsets of variables, as measured by the diverse scintillation indices, (iii) the identification of representative variables, and (iv) the interplay of variable behaviors at different temporal scales. Moreover, it was clear from the start that our collaborators were interested not only in an overview of variables behavior, but also required to identify values at specific time points along data exploration. Hence, it is important to capture multivariate nature and temporal behavior while preserving detailed information, with enough display space to visualize details over time.

Many data mining and visualization contributions address the identification of relevant variables, analysis of spaces of variables, or the investigation of the behavior of individual variables or groups of variables, as discussed in Chapter 2. Solutions focused on the analysis of variables are associated with visual representations which (i) are not generic, (ii) do not allow incorporating user knowledge or views, or (iii) do not suffice to answer all the questions identified associated with ionospheric data analysis. Hence, our goal is to propose and validate new approaches that consider the ionospheric data requirements and incorporate interaction mechanisms to support diverse user-driven analysis tasks.

Our *TV-MV Analytics* visual framework combines multiples visualizations that convey the temporal behavior of individual target variables with clustering and multidimensional projection techniques in order to (i) provide data analysts with a visual exploration loop that supports inspecting the multivariate values recorded, at different temporal scales; (ii) provide hierarchical visualizations that convey the temporal behavior of individual variables and also of groups of related variables; (iii) uncover how the behaviors of multiple variables are related, by means of a small multiples visualization associated with clustering and multidimensional projection techniques; (iv) allow interactive exploration of the multiple variable spaces as descriptors of the data instances; and (v) allow a user to assess how different variable spaces characterize an observed behavior.

In this chapter, we provide an overview of *TV-MV Analytics*, detailing resources available for a possible data exploration pipeline (Section 4.2), the software architecture (Section 4.3), the data retrieval functions (Section 4.4), the data preparation and processing, data analytics and derived representations (Section 4.5), the visualizations of data variables (Section 4.6), and the visualization of data instances (Section 4.7).

4.2 Overview of *TV-MV Analytics* Framework

An overview of the data processing and exploration stages is shown in Figure 27, which is divided in ten regions. The main interface menu corresponds to the region labeled A, which contains the links to access the different functionalities. The user must initially specify the data to be retrieved: the subset of variables, the data sources, the target time period, the data aggregation function (recorded values may be aggregated) and temporal granularity (analysis time unit, e.g., hour, day, etc.) (region B). The query returns as the data set for analysis the observations described by the aggregated values of every selected variable along the selected time period.

The following steps are concerned with data processing (regions C and D). Variable's aggregated values are linearly normalized to the interval $[0,1]$, and missing values may be interpolated. A time matrix representation of each variable is created after processing, where each entry stores a single normalized value (region C) relative to the user-defined time granularity, which provides the basis of all visualizations and analytics. From each time matrix representation we derive a multidimensional feature vector, which is employed in the subsequent mining and visualization stages. The feature vectors are clustered choosing from the X-means, K-means, Bisecting K-means, K-medoids, or SC optimized-K-means clustering algorithms (region D), to identify groups of variables showing similar behavior over the selected time period and time scale.

Finally, visualizations of both variables and data instances are created. The visualizations of variables include (i) a hierarchical similarity map of their feature vectors (region F) where

Figure 27 – Overview of resources in *TV-MV Analytics*. (A) Main menu displays major functionalities. (B) Query panel for retrieving observations described by a subset of user-selected variables (16 selected in this example) over a time range. (C) Derived time matrix representation of each variable (after normalization/interpolation); (D) Derived feature vector representation of time matrices, feature vectors clustered. ((C) and (D) refer to data processing steps). (E) Small multiples time matrix view of all variables. (F) Similarity map view of all variables. (G) Interface to modify cluster model or projection technique. ((E) and (F) refer to multiple coordinated visualizations of both spaces of variables and data instances). (H), (I) Timeline views of user-selected variables. (J) Data analytics: classification, regression and prediction functions.

each colored circle represents a variable and colors identify its cluster, (ii) time matrices views for the variables (region E) where header colors map their corresponding cluster, (iii) correlation matrices, (iv) and time circular diagram (not shown), which convey further insight into their temporal behavior and possible correlations. Region G shows information on the quality of the cluster model and functions to modify it, or to select another projection technique for the similarity map.

Region H indicates that four variables have been user-selected to be shown as aggregated and normalized line-graphs (region I) to convey seasonal patterns and hierarchical similarity-based visualizations of the user-defined data observations. Views can be inspected and manipulated and the current cluster model refined to reflect user knowledge and beliefs. Analysis of these visualizations can be combined with results from data mining techniques such as classification (J), regression and prediction algorithms, so experts can assess the role of alternative variable spaces as representative descriptors of the target variable.

4.3 Software Architecture

We adopted the multi-tier architecture (SCHULDT, 2009), with the software components organized into: i) a data tier, ii) a business logic tier, iii) a presentation tier, and an iv) tier with dedicated functionality for data access. The architecture and its components is presented in Figure 28, and a brief description follows:

1. In the data tier, a PostgreSQL server provides data storage, access functions, and storage procedures to ensure real time data exploration. A relational PostgreSQL database was thus created for each data set considered, with an underlying common general data model as shown in Figure 29. Basically, it consists of tables *source1*, *instance* and *variable*. Table *variable* contains the identifier, the name and the category of each variable. Table *source1* describes the sources from which the data can be retrieved, such as satellites for the ionospheric scintillation data. Table *instance* contains the identifier and the values for each variable considered, source and time. Functions and storage procedures were implemented in the PL/SQL procedural language to retrieve aggregated values, using function *generate_subscripts* to speed-up the process.
2. An XML configuration file was defined in the data access tier (see Appendix B) to connect the framework with multiple databases originating from distinct domains. It defines connection data, table names and elements, PL/SQL functions names and parameters, and other configuration parameters. These parameters define the exploration options and will be further described later. The configuration file allows adjusting the system to handle generic time-varying multivariate data sets.

Figure 28 – Software Architecture.

Figure 29 – General database model.

3. The business logic tier combines Java components with external sources (highlighted in yellow) and Java servlets to handle clients's requests and return custom responses, including visualizations. The external sources include Weka libraries for classification and regression algorithms and in-house components (developed at VICG) that implement projection techniques and clustering methods.
4. Finally, the presentation tier provides the user interface including all visualizations and interaction functions. In order to provide user-friendly views we combined JavaServer pages, HTML, Cascading Style Sheets (CSS), Bootstrap, Ajax, JSON and some Javascript libraries such as D3.js, JQuery, JodaTime and RickShaw.

4.4 Query Window

A time-varying multivariate data set is described by observations of n variables $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n\}$, where each variable $\mathbf{x}_i$ is a sequence of time-stamped numerical values recorded at a particular temporal rate (minute, hour, day, month or year), i.e.,

$$\mathbf{x}_i = \{x_i^{t_1}, x_i^{t_2}, x_i^{t_3}, \dots, x_i^{t_k}\}. \quad (4.1)$$

We define a data instance at a time t_j as the vector $\mathbf{o}^{(t_j)}$, with the values of p selected variables ($p \leq n$), taken at a time t_j , i.e.,

$$\mathbf{o}^{(t_j)} = [x_1^{t_j}, x_2^{t_j}, x_3^{t_j}, \dots, x_p^{t_j}]. \quad (4.2)$$

Finally, we define aggregated data instances as vectors $\mathbf{o}^{(t_{se})}$, which are obtained by applying any aggregation function (e.g., maximum, minimum, average, median or standard deviation) to the retrieved data instances over a time interval $[t_s, t_e]$ of size $l = e - s$, and can also be defined as a vector of p values, i.e.,

$$\mathbf{o}^{(t_{se})} = [x_1^{t_{se}}, x_2^{t_{se}}, \dots, x_p^{t_{se}}], \quad (4.3)$$

where $x_i^{t_{se}} \in \mathbf{o}^{(t_{se})}$ is the result of applying the selected aggregation function $f : \mathbb{R}^l \rightarrow \mathbb{R}$ to a subset of values of the vector $\mathbf{x}_i$ over a time interval $[t_s, t_e]$, i.e.,

$$x_i^{t_{se}} = f(x_i^{t_s}, x_i^{t_{s+1}}, x_i^{t_{s+2}}, \dots, x_i^{t_e}). \quad (4.4)$$

The Query Screen is the initial window of *TV-MV Analytics* which is guided by the defined XML configuration file to retrieve information from a PostgreSQL database (see Appendix B) and is divided into four regions (A, B, C and D). An example is shown in Figure 30. Depending on user input, the query may produce a data set of instances of type $\mathbf{o}^{(t)}$ or $\mathbf{o}^{(t_{se})}$.

- In **Region A** in Figure 30, it is possible to select the subset of target variables, individually, or as category groups. A label shows the number of selected variables at the bottom of the variables list.
- **Region B** allows to define the temporal range, aggregation function and analysis time unit. The choice of time units is domain and data-dependent, and possibilities must have been previously informed in the XML file. The aggregation functions allowed are shown, which can be the maximum, the average, the minimum, the median or the standard deviation of the retrieved values, so that database values will be aggregated over the defined time-unit interval accordingly.
- If any threshold is defined to filter data instances from a particular database variable in the XML configuration file, the variable and available options are shown in **Region C**. For the ionospheric scintillation data set it shows the minimum elevation threshold, set up to avoid retrieving data affected by the so-called multipath effects (JIAO *et al.*, 2013). Additional options included for variables *source1* (in the case of scintillation, it informs the satellite from which data will be retrieved), and *source2*, an optional input to inform from which station data will be retrieved.
- **Region D** shows the choices of matrix moments (described in Section 4.5.2) that may be employed as feature vector descriptors of the time matrices.

Given the input information, a query is built and processed, if it is successful the following message is shown: “*Query was successfully processed*”. Otherwise, if data is not available over the queried period, it shows: “*Query without result*”. Then, the query produces a set of aggregated data instances comprising a subset of the time-varying multivariate data available in the database.

4.5 Data Processing

Usually the multiple variables take values on distinct units and orders of magnitude and must undergo some normalization prior to further processing. Consequently, data ranges may be linearly normalized to the interval $[0, 1]$, according to Equation 4.5.

$$x_{norm} = \frac{x - min}{max - min}, \quad (4.5)$$

where x is the original value, max and min the maximum and minimum value of the data subset, respectively.

Collected data may include many missing values, which can either characterize typical behaviors or measurement errors. Our framework adopts one of three strategies to handle this problem:

The interface is divided into four sections marked with colored circles and dashed lines:

- A (Green dashed box):** A list of attributes under the heading "Attributes:". It includes a "Select All Attributes" checkbox and two groups of attributes: "direction_indicator" (azim, elev) and "index" (a long list of variables like avgccd, avg_cn, dtec, f2nd_tec, numerator, phi, etc.).
- B (Pink dashed box):** "Date Range:" with "From:" and "To:" input fields, and "Time Unit:" with a dropdown menu set to "minute(s)".
- C (Blue dashed box):** "Aggregation Function:" with a dropdown menu set to "max", and "Minimum elev:" with a dropdown menu set to "0".
- D (Orange dashed box):** "station:" with a dropdown menu set to "PRU1", "satellite:" with a dropdown menu, and "Moments:" with a dropdown menu set to "Statistical Moments - Invaria".

A "Visualize" button is located below the Moments section. A large empty rectangular area is at the bottom of the window.

Figure 30 – Query window: the initial window of *TV-MV Analytics* is divided into four sections for selecting: (A) a subset of target variables, (B) the temporal range, aggregation function and analysis time unit, (C) the threshold options and source, and (D) matrix moments.

1. If a value at a particular time instant is missing, but values are known for its previous and subsequent time points, the missing value will be obtained from linear interpolation of neighboring values;
2. If values for a whole day are missing, but values are known for its previous and subsequent days, the missing values of the entire day will be obtained from linear interpolation of neighboring day values;
3. If the previous cases do not apply, values will not be estimated and the respective entries remain unknown.

4.5.1 Time Matrix Representation

A time matrix representation summarizes the temporal behavior of a particular variable for the user-defined observed period. As the user may choose to aggregate the observed values, a

time matrix maps the values of a variable along a certain period and at a given time granularity, from minutes to hours or even days. Each matrix entry represents a single numerical value, either as recorded in the database or computed with Equation 4.4, by applying an aggregation function over a range of values.

The user selects the initial date d_s and the final date d_e of the target period in milliseconds, the number of time units $\#_{AG} = \{\#_{AG} \in \mathbb{I} | \#_{AG} > 0\}$ and the aggregation time unit tu_{AG} (converted in milliseconds), which define the temporal granularity, which may be seconds to months or even years. Also, a row time unit tu_R (converted in milliseconds) is selected, which defines the primitive for each row in the time matrix.

The time matrix $\mathbf{M}_{n \times m}$, in which $n = (d_s - d_e)/tu_R$ and $m = \lceil \times \rceil \lceil \times \rceil tu_R/tu_{AG}/\#_{AG}$, is formed by entries that represent the values resulting from applying the selected aggregation function, over the defined time unit, to the variable's recorded values in the database. That is, from the values of the variable $\mathbf{x}_i$, defined in Equation 4.1, we applied the selected aggregation function f to a subset of values over a certain time interval according to the position of each entry, as defined in Algorithm 1.

Algorithm 1 – Time Matrix Representation

```

1: procedure TIME MATRIX CONSTRUCTION( $\mathbf{x}_i, d_s, d_e, \#_{AG}, tu_{AG}, tu_R$ )
2:    $\mathbf{x}_i$ ;                                ▷ values of variable  $i$ ;
3:    $d_s$ ;                                ▷ initial date;
4:    $d_e$ ;                                ▷ final date;
5:    $\#_{AG}$ ;                                ▷ number of units;
6:    $tu_{AG}$ ;                              ▷ aggregation time unit;
7:    $tu_R$ ;                                ▷ row time unit;
8:    $n \leftarrow (d_s - d_e)/tu_R$ ;
9:    $m \leftarrow \lceil tu_R/tu_{AG} \rceil / \#_{AG}$ ;
10:  for  $a \leftarrow 1$  to  $n$  do
11:    for  $b \leftarrow 1$  to  $m$  do
12:       $index \leftarrow a \times m + b$ ;          ▷ index position
13:       $d'_s \leftarrow d_s + index \times (\#_{AG} \times tu_{AG})$ ;    ▷ initial date of period
14:       $d'_e \leftarrow d'_s + (\#_{AG} \times tu_{AG})$ ;          ▷ final date of period
15:       $\mathbf{x}'_i \leftarrow \mathbf{x}_i [d'_s \rightarrow d'_e]$ ;          ▷ subset of values of  $\mathbf{x}_i$ 
16:       $M_{ab} \leftarrow f(\mathbf{x}'_i)$ ;                      ▷  $f$  is the aggr. function
17:    end for
18:  end for
19:  return  $\mathbf{M}_{i_{n \times m}}$ ;                      ▷ time matrix for variable  $i$ 
20: end procedure

```

4.5.2 Feature Moments

Typical multivariate data analysis approaches operate directly on the set of data instances. In the case of time-varying multivariate data, this is not feasible because (a) there is a high

number of data instances, (b) it is important to consider the temporal sequence of the data instances and, in this case, (c) there is a need to explore the behavior of the variables. Thus, we use feature vectors from the time matrix representations in order to capture the characteristics of the variable behavior over time.

TV-MV Analytics offers multiple choices to compute feature vectors representative of the time matrices structure, such as Statistical Moments (also known as Color Moments), Hu moments, Normalized Central Moments, Abo-Zaid Moments, and Orthogonal Variant Moments, summarized in Table 15 and described next.

Table 15 – Different choices of feature vectors to represent the internal time matrices. # represents the size of the resulting feature vectors.

Moments	Reference	#	Description
Statistical (Color)	Tian <i>et al.</i> (2013)	5	invariant to scale and rotation
Hu	Hu (1962)	7	invariant to translation, scale and rotation
Normalized Central	Maitra (1979)	16	invariant to translation, and scale
Abo-Zaid	Abo-Zaid, Hinton and Horne (1988)	16	invariant to position, contrast, and size, regarding to intensity variations and scale changes
Orthogonal Variant Moments (OVM)	Santos, Lope <i>et al.</i> (2010)	5	orthogonal variants, for sensitive to perturbation

Color moments allow analyzing similarity between images considering a probabilistic distribution of its colors. These features are scalable and invariant to image rotation due to their statistical nature. [Tian *et al.* \(2013\)](#) consider three main color moments: Mean, Standard Deviation and Skewness (the symmetry of the color distribution). Besides, we consider the Kurtosis moment, that captures the proportion of variance regarding the concentration of near and far values to the Mean. We also added a uniformity moment, which has a high value if all colors are equal or constant. The color moments feature vector is formed by the five standard statistical moments computed from the internal matrix, i.e., mean, standard deviation, skewness, kurtosis, and uniformity, defined in Table 16. They capture the properties of the probability distributions that characterize the variables' behavior over the observed time period. Matrix entries with missing values are ignored in the computation of statistical moments. However, it is possible to include the percentage of missing data as an extra feature in the vector, for situations in which such a variable should be clearly identified as an outlier.

Hu moments have desirable characteristics, i.e. they are orthogonal invariants and independent of position, size and orientation. They capture the shape of the variable's distribution over the observed period. They are computed from normalized moments up to order three, with

Table 16 – Statistical Moments (Color Moments).

Mean	$\bar{x} = \frac{1}{N} \sum_{j=1}^N x_j$
Standard deviation	$\sigma = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2}$
Skewness	$\gamma_1 = \frac{1}{N} \sum_{j=1}^N \left(\frac{x_j - \bar{x}}{\sigma} \right)^3$
Kurtosis	$kurt = \frac{1}{N} \sum_{j=1}^N \left(\frac{x_j - \bar{x}}{\sigma} \right)^4$
Uniformity	$unif = 1 - (\sigma/\bar{x})$

the equations described in Table 17:

Table 17 – Hu moments (HU, 1962).

$\mathbf{H}_1 = \eta_{20} + \eta_{02}$
$\mathbf{H}_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$
$\mathbf{H}_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$
$\mathbf{H}_4 = (\eta_{30} - \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$
$\mathbf{H}_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right]$ $+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$
$\mathbf{H}_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$ $+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$
$\mathbf{H}_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$ $- (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

In Table 17, $\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}$, $\gamma = (p + q + 2)/2$ is the normalized central moment, p and q are the non-negative integer degrees of the moment, μ_{pq} the central moment, x_j a matrix value, and N the number of matrix entries. The first moment is the inertia around the matrix's centroid, where the intensities are analogous to physical density. The first six moments encode a shape with invariance to translation, scaling and rotation. The seventh moment ensures skew invariance, which enables to distinguish between mirrored matrices of otherwise identical matrices.

The normalized central moments of a matrix $M(i, j)$, where i, j are the coordinate axes, are based in the classical central moments. An important concept is the center of gravity $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$, where m_{pq} is the $(p + q)^{th}$ order moment, defined in Equation 4.6.

$$m_{pq} = \sum_i \sum_j i^p j^q M(i, j). \quad (4.6)$$

The central moments are defined about that center of gravity, given by Equation 4.7.

$$\mu_{pq} = \sum_i \sum_j (i - \bar{x})^p (j - \bar{y})^q M(i, j), \quad (4.7)$$

where $p \in [0 - 3]$, $q \in [0 - 3]$ and μ_{pq} is the $(p + q)^{th}$ order central moment. These moments are invariant to translation and, once normalized, also to scaling. The normalized central moments

are given by Equation 4.8.

$$n_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}, \quad (4.8)$$

where $\gamma = \frac{(p+q+2)}{2}$.

Abo-Zaid, Hinton and Horne (1988) suggest a variation of Hu's moments, defined by a new moment normalization to cancel the scale and contrast changes before the computation of the moment invariants. The variation normalization factor is defined in Equation 4.9. Authors claim this modification turns the moments position, contrast, and size invariant regarding to intensity variations and scale changes.

$$\mu'_{pq} = \mu_{pq} \frac{1}{\mu_{00}} \left(\frac{\mu_{00}}{\mu_{20} + \mu_{02}} \right)^{\frac{p+q}{2}} \quad (4.9)$$

Finally, we include an Orthogonal Variant Moments (OVM) option introduced by Santos, Lope *et al.* (2010). These moments, on the contrary, are designed to be sensitive to perturbation, to measure a transformation, and are orthogonal (i.e. uncorrelated). They include a normalization factor η on the total matrix size ($height \times width$), defined in Equation 4.10.

$$\eta = \frac{1}{height \times width} \quad (4.10)$$

The first moment is the area, defined by $A = n \sum_y \sum_x M(i, j)$. The moments Lx and Ly are two orthogonal components to calculate the surface area (Equation 4.11 and 4.12 respectively).

$$Lx = \eta \left[1 + \sum_y 1 + \sum_x \sqrt{1 + \left(\frac{\partial M(i, j)}{\partial x} \right)^2} \right] \quad (4.11)$$

$$Ly = \eta \left[1 + \sum_x 1 + \sum_y \sqrt{1 + \left(\frac{\partial M(i, j)}{\partial y} \right)^2} \right] \quad (4.12)$$

The last two moments Dx and Dy are given to calculate the position or stations, a similar approach to the traditional center of mass, and are defined by Equations 4.13 and 4.14.

$$Dx = \eta \sum_x \sum_y (x+1) M(i, j) \quad (4.13)$$

$$Dy = \eta \sum_x \sum_y (y+1) M(i, j) \quad (4.14)$$

4.5.3 Data Analytics

In order to group variables, feature vectors as descriptors of the time matrices are clustered based on global similarity of their behavior over the observed time period. *TV-MV Analytics* offers a choice between K-means, Bisecting K-means, K-medoids and X-means algorithms¹, plus an SC optimized-K-means clustering algorithm. It also offers a choice between Euclidean distance or Inverse Pearson Correlation (simulating dissimilarity, described in Equation 4.15) as dissimilarity measures.

$$\rho(\mathbf{x}_i^{\text{tse}}, \mathbf{x}_j^{\text{tse}}) = 1 - \left| \frac{\text{cov}(\mathbf{x}_i^{\text{tse}}, \mathbf{x}_j^{\text{tse}})}{\sigma_{\mathbf{x}_i^{\text{tse}}} \sigma_{\mathbf{x}_j^{\text{tse}}}} \right|, \quad (4.15)$$

in which $\mathbf{x}_i^{\text{tse}}$ is an aggregated variable, cov is covariance, σ standard deviation and $||$ to consider either positive or negative correlation coefficients.

K-means departs from a choice of k data points as tentative initial cluster centroids (MACQUEEN *et al.*, 1967). Then, each data element is associated with its closest centroid, forming clusters. The cluster centroids are then updated based on the elements actually assigned to the cluster. The assignment is repeated until no changes are observed in the clusters. The K-medoids algorithm follows the same rationale, but using the cluster medoids, the most representative cluster element (KAUFMAN; ROUSSEEUW, 1987). Both algorithms are sensitive to the initial selection and require the number of desired clusters as input.

In the bisecting K-means algorithm the rationale is to start splitting the data set into two clusters, selecting one of these to split further and so on, until obtaining a desired number of k clusters (STEINBACH *et al.*, 2000). There are different ways to choose which cluster to split. A popular choice is to select the one with the highest sum of squared error (SSE), defined in Equation 4.16, which is the total sum of the distance from each element to the closest centroid.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} \|c_i - \mathbf{x}\|^2, \quad (4.16)$$

where k the number of clusters, C_i is the i^{th} cluster, and c_i the centroid of cluster C_i .

The X-means algorithm is a hierarchical iterative implementation of K-means that does not need k be known *a priori*, as it learns this number as the one that yields the best Bayesian Information Criterion (Equation 4.17) (PELLEG; MOORE *et al.*, 2000).

$$BIC = n \times \ln(\hat{\sigma}_e^2) + k \times \ln(n), \quad (4.17)$$

where n is the number of data elements, k the number of free parameters to be estimated, and $\hat{\sigma}_e^2$ is the error variance defined as $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

¹ Clustering algorithm implementations are available as *Vispipeline* components (PAIVA *et al.*, 2011).

We introduce yet another variation of K-means that selects k as the value that yields the best Silhouette Coefficient (SC) (ROUSSEEUW, 1987). The silhouette coefficient of a single data point is defined as the relation between its average distance to points in its cluster and its minimum average distance to points in another cluster (Equation 4.18).

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}, \quad (4.18)$$

where a_i is the average distance from a i^{th} data point to all the other points in its own cluster and b_i the average distance of that data point to all the points in the other clusters.

An overall measure of the goodness of the clustering model is the average silhouette coefficient of all points. Silhouette values are in the range $[-1, +1]$, where higher values indicate a cluster model with higher cohesion and separability. Negative values are undesirable because they indicate the average distance between points within clusters is greater than the minimum average distance to external points.

TV-MV Analytics allows users to re-assign variables to clusters or to create new clusters based on their consideration of element similarity, e.g., after inspection of time matrix views and similarity maps of variables, as described in the next Section. Flexibility in defining the cluster model is mandatory in situations where analysts need to compare the behavior of alternative feature spaces. Indeed, the analyst's perception of relevance and similarity of multiple variables may be different from the one inferred by an automatic clustering. *TV-MV Analytics* encourages investigating alternatives, where users' knowledge is fundamental to define the assignment of variables to clusters than silhouette or other measurements. In any case, the reference SC values are immediately updated whenever a current cluster model is modified.

Each resulting cluster is uniquely associated with a random color, following the colormap² shown in Figure 31, which provides color definitions for a very extensive range of colors (KUMAR, 2016). To ensure maximum color differentiation we empirically determine those colors at a RGB Euclidean distance above 100 to all the others, as detailed in Algorithm 2.

Figure 31 – Colormap used to characterize each resulting cluster.

Beyond the clustering algorithms to identify groups of variables with similar behavior along the observed period, we design *TV-MV Analytics* to incorporate strategies to identify representative spaces and to evaluate them through classification and regression functions. Departing from a clustering of the variables a user analyst may define a representative space of variables choosing from one of five strategies:

² Colormap available at <http://latexcolor.com/>.

Algorithm 2 – Color mapping for each resulting cluster.

```

1: procedure COLOR MAPPING( $n, \mathbf{c}$ )
2:    $n$ ; ▷ number of clusters;
3:    $\mathbf{c}$ ; ▷ resulting color vector;
4:   if size( $\mathbf{c}$ ) ==  $n$  then
5:     return  $\mathbf{c}$ ;
6:   end if
7:   if size( $\mathbf{c}$ ) == 0 then
8:      $\mathbf{c.add(rand(255))}$ ;
9:   else
10:     $r = rand(255)$ ;
11:    for  $i \leftarrow 1$  to size( $\mathbf{c}$ ) do
12:      if  $\|\mathbf{c}(i) - r\| \leq 100$  then
13:        go to 10;
14:      end if
15:    end for
16:     $\mathbf{c.add}(r)$ ;
17:  end if
18:  COLOR MAPPING( $n, \mathbf{c}$ );
19: end procedure

```

Figure 32 – Five strategies for variable space selection: (a) cluster medoids, (b) user choice, (c) k variables closest to cluster centroid, (d) result of simple linear regression, or (e) relevant variables from linear regression function. Dashed gray circles in (a) and (c) represent cluster virtual centroids. Red circle in (a) is the medoid and red circles in (c) are points closest to the virtual centroid. Yellow circles in (b) represent user-selections. Green circles in (d) and (e) represent linear regression results. Blue circles represent unconsidered variables.

- **Cluster medoids.** One variable from each cluster is automatically picked as representative, which will be the one associated with element closest to the cluster centroid (Figure 32.a).
- **User-selection.** Analysts can modify the automatic strategy selecting any alternative variables as a cluster representatives (Figure 32.b).
- **k variables closest to the centroid.** Cluster representatives are selected as the k variables closest to the cluster centroids, for a user-input k .
- **Simple linear regression.** In this strategy the goal is to pick the variable most correlated to a user-defined reference variable as cluster representative. A simple linear regression

model (Equation 4.19) is learned, and the cluster representative is picked as the variable that yields the lowest squared error (ALENCAR *et al.*, 2017).

$$y = \beta_0 + \beta_1 \times x, \quad (4.19)$$

where y is the response, β_0 and β_1 are the y intercept and slope, respectively, of the regression line, and x is the reference variable.

- **Results of linear regression.** The rationale is similar to the previous case, but a linear regression model is learned considering all variables (Equation 4.20) instead. The Akaike criterion (AIC) for model selection is employed to decide which variables to be included in the regression model (Equation 4.21). Any variables not included in the regression model are discarded from the set of representative variables.

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_n \times x_n, \quad (4.20)$$

where y is the response, β_0 the y intercept of the regression line, β_i is a parameter to generate the linear curve, and x is the variable.

$$AIC = 2 \times k - 2 \times \ln(\hat{L}), \quad (4.21)$$

where $\hat{L}$ is the maximum value of the likelihood for the regression model and k the number of variables.

TV-MV Analytics includes implementations of decision tree and multilayer perceptron classification algorithms, as well as of linear regression and simple linear regression algorithms³. Analysts may execute them to obtain further information regarding the alternative variable spaces considered and assess potential data descriptors effective for characterizing a target variable. Decision trees and neural networks (such as the multilayer perceptron, which is a feed-forward neural network) have been chosen because they are widely known and have been successfully applied in many pattern classification applications (KOH; TAN *et al.*, 2011), although other algorithms might be incorporated into the framework.

4.6 Visualization of Variables

In this section we describe two-dimensional (2D) visual representations for depicting variable behaviors over time. They are aimed at visualizing multivariate data sets with many variables and a large number of data instances. These visualizations support different tasks,

³ Regression algorithms were imported from the Weka libraries (FRANK *et al.*, 2009) and used in the case studies with the default Weka parameters.

such as identifying relationships between variables, comparing their behaviors, similarity or correlation over different periods, analyze the relevance of variables to characterize a behavior, identify unusual data patterns, occurrence and selectivity of missing values, and analyzing variable spaces at different temporal scales.

4.6.1 Small Multiples Time Matrix Visualizations

Each time matrix representation has a corresponding visualization (Figures 33 and 34) which is split into two areas. The header area shows the variable's name, its assigned cluster (also indicated by the header and border color), the source from which data instances have been retrieved, the data instance period, the minimum and maximum values recorded over that period and the rate (percentage over total) of missing data before and after interpolation.

Figure 33 – Time matrix visualization of variable *phi0111*. Daily data (00:00h-23:00h) from GPS satellite 1 at station PRU2, from June 1 to 30, 2014. In this example, cells depict the maximum value observed over a 30-minute interval (user choice of aggregation function and time span). Each row depicts values for a day and each column shows values for the same thirty-minute period, on different days. Areas in gray depict matrix cells with missing values, and cells enclosed by a red border show interpolated data. The matrix border in cyan indicates a high rate of missing data.

The main area is split into cells depicting corresponding matrix entries. Values are color-mapped using the heated-object colormap (LEVKOWITZ, 1997), where brighter colors map to lower values and darker colors map to higher ones (Figure 35).

Figure 34 – Another time matrix visualization of the same variable *phi0111*, now depicting monthly data from January 1 to December 31, 2014. In this example, each cell depicts the median value observed over a day. Each row shows daily values observed along a specific month and each column shows values relative to the same sequential day along the 12 months. Regions marked A and B show the detail labels overlaid upon mouse hovering over the indicated cells (shown separately for clarity).

Figure 35 – Heated-Object colormap (LEVKOWITZ, 1997).

Cells that show interpolated data are highlighted by a red border, and cells depicting unknown entries (missing data) are shown in gray without a border. The border of a time matrix may be highlighted in cyan to indicate it has a missing data rate above a user-defined threshold (say, above 50%). In our particular case, the choice of colors is based on recommended red and cyan colors to catch viewer attention (SAMARA, 2007). They are also complementary hues (LEVKOWITZ, 1997).

In the examples, cells show the maximum value observed within a 30-minute interval (Figure 33) and the median value observed over a day (Figure 34). Hovering over a cell causes a label to show up with detailed information about the cell's corresponding entry: date, recorded value, and aggregation function and time span. This is illustrated for the two marked regions A and B of the time matrix depicted in Figure 34.

Time matrix visualizations shown in Figures 33 and 34 can be viewed individually or as a small multiples visualization of all queried variables, where views are spatially arranged according to their assigned clusters. There are three possible layouts to arrange the small multiples: (1) by cluster, (2) by source, or (3) alphabetically under the variable's name. A combo

box, shown in the header area of each matrix view, allows analysts to modify the corresponding variable's cluster assignment.

Moreover, the time matrix representation of variables can assist in prediction tasks, based on the historical data. A functionality is included to predict the series of values in the next-row in a given time matrix. This is achieved by applying Isotonic regression functions (WITTEN *et al.*, 2016), to the time matrix columns, as illustrated in Figure 36. The correlation between the series of predicted and actual values can be compared.

The Isotonic regression function can be viewed as a least squares problem under order restriction. It is defined as a monotonic function best fitting to the original data instances that minimizes Equation 4.22:

$$f(x) = \sum_{i=1}^n w_i (y_i - x_i)^2, \quad (4.22)$$

where $Y = (y_1, y_2, \dots, y_n)$ is a set of real numbers representing the observed values and $X = (x_1, x_2, \dots, x_n)$ are the unknown values, according to the order restriction: $x_1 \leq x_2 \leq \dots \leq x_n$, and w_i are positive weights.

Figure 36 – Example of prediction task using Isotonic Regression. An isotonic function is created for each column. The series values in the next-row are predicted using that function based on historical data.

4.5.1.1 Hierarchical Small Multiples Time Matrix Visualizations

Time matrices views tend to take up a lot of space, which is aggravated when the matrix describes an extended time period. This creates a disadvantage in scenarios where many variables are analyzed simultaneously: (i) it is difficult to display the small multiples, and (ii) hard for the analyst to inspect many time matrices simultaneously. Hence, we propose a hierarchical version of the small multiples view. It relies on the cluster tree resulting from the X-means algorithm, illustrated in Figure 37.

Figure 37 – Schematic visualization of the hierarchical time matrices. Circles represent feature vector of time matrices, which are the input data for the clustering tree. The root contains the medoids of each input cluster, each level contains submedoids, and leaves store each feature vector point.

Leaf nodes represent single variables, whereas nodes up in the hierarchy summarize the contents of a cluster by showing the time matrix of its medoid element. Each cluster is sub-clustered with X-means, defining new sub-cluster nodes displayed as their corresponding medoids, repetitively until reaching the leaf nodes. Algorithms 3 and 4 describe the process of generating the hierarchical tree representation of the variables, which can be explored at multiple levels. The hierarchical tree representation is coordinated with the hierarchical similarity map, described in Section 4.6.3.

Algorithm 3 – Clustering Node and Clustering Tree

```

1: Struct ClusteringNode {
2:   Int depth;                                ▷ depth of the node
3:   Int members[sn];                          ▷ sn indices of cluster members
4:   Int medoids[sm];                          ▷ sm indices of cluster medoids
5:   ClusteringNode* children[sm];             ▷ pointers to the sm sub-clusters
6:   ClusteringNode parent;                   ▷ pointer to the parent
7: };
8: Struct ClusteringTree {
9:   Int depth;                                ▷ total tree depth
10:  ClusteringNode root;                      ▷ pointer to the root
11: };

```

Algorithm 4 – Clustering Tree Construction

```

1: nodeParent  $\leftarrow$  Initial nodeParent = root;
2: cM  $\leftarrow$  Clustering model;
3: function CLUSTERING TREE( cM, nodeParent )
4:   newCM  $\leftarrow$  X – means(cM);
5:   if cM.medoids.size == newCM.medoids.size then
6:     newNode  $\leftarrow$  new ClusteringNode(cM.depth+1, cM.members, cM.medoids,
7:     nodeParent);
8:     return Leaf(newNode);
9:   else
10:    for i  $\leftarrow$  1 to newCM.clusters.size do
11:      newNode  $\leftarrow$  new ClusteringNode(cM.depth+1, newCM[i].members,
12:      newCM[i].medoids, nodeParent);
13:      nodeParent.children[i]  $\leftarrow$  newNode;
14:      return CLUSTERING TREE(newNode, nodeParent.children[i]);
15:    end for
16:  end if
17: end function

```

4.6.2 Time Circular Diagram

The time matrices are capable of conveying a variables' global behaviors while preserving detail. However, analysts often are interested in inspecting whether a certain variable displays correlations within along time intervals within the observed period. This motivated us to introduce the time circular diagram visualization. This visualization is also split into two areas (Figure 38), a header area with information on the variable and a main area, as in the time matrix representation. The difference is that the main area displays the normalized entries of the time

Figure 38 – Time Circular Diagram of a variable in the air quality data set, showing the correlations of 1-day periods from March 10-31, 2004. The sub-periods are placed clockwise, where the initial day (March 10) is placed at the middle top.

matrix $\mathbf{M}$, now placed as cells in a circular clockwise distribution, and we also adopt the heated-object colormap (LEVKOWITZ, 1997).

The analyst can define the target temporal window, or sub-period, to assess possible correlations. Lines connect sub-periods if their correlation is above a user-defined threshold (0.8 as default, which means very high correlation (TODOROVA; VASSILEV; SURCHEV, 2016)). An edge bundling strategy (PUPYREV; NACHMANSON; KAUFMANN, 2010) is applied to the connecting lines to reduce clutter, grouping similar edges in bundles. Sub-periods not connected by lines indicate time windows with no correlation above the threshold. The sub-periods can be reduced by using the Haar wavelet (LEPIK; HEIN, 2014) to simplify the time series representation and produce a roughly approximation at a lowest resolution level. The maximal resolution *level* is calculated using the sub-period size *sp_size*, defined in Equation 4.23. This feature is useful to explore data relative to extensive temporal intervals.

Similarly to the time matrices views, time circular diagrams can be viewed individually or as a small multiples visualization of all queried variables arranged in a layout according to the clustering model.

$$\begin{aligned} sp_size &= 2^{level} \\ level &= \log_2 sp_size \end{aligned} \quad (4.23)$$

in which sp_size is the sub-period size defined by the user and $level$ is the resulting maximal resolution level.

4.6.3 Similarity Map of Variables

Similarity maps assist users to handle high dimensional data for visualization in lower dimensional spaces taking advantage of human visual ability to recognize patterns, such as clusters and neighborhoods (PAULOVICH; OLIVEIRA; MINGHIM, 2007). A similarity map depicts a two-dimensional space embedding of the m -dimensional feature vectors describing the variables' time matrices (as described in Section 4.5.2), in which 2D point distances are used as proxies to the point distances in the original m -dimensional space (TEJADA; MINGHIM; NONATO, 2003; MINGHIM; PAULOVICH; LOPES, 2006). Therefore, “more similar” variables (as described by their corresponding time matrix) are placed closer in the 2D map, whereas more “dissimilar” ones are placed farther apart.

In *TV-MV analytics* these maps are aimed at assist specialists in assessing, and modifying, the clustering results and indication of representative variables according to their appreciation of similarity in variable behavior. The 2D embedding is computed with a multidimensional projection called IDMAP (MINGHIM; PAULOVICH; LOPES, 2006). IDMAP combines two techniques: (i) it generates an initial mapping with Fastmap (FALOUTSOS; LIN, 1995), and (ii) iteratively improves point placement using the Force Scheme strategy (TEJADA; MINGHIM; NONATO, 2003), recovering information lost in the initial projection. Pairwise Euclidean or the Inverse Pearson correlation may be considered to compute the dissimilarities between feature vectors. Analogous to the time matrix visualization, circles can be highlighted with a cyan border to indicate missing data above a threshold (50%). Each colored circle depicts a variable and its associated cluster, preserving the same color coding of the clusters adopted in the small multiples matrix view.

The map conveys the overall similarity between variables and their implicit groupings, regardless of their explicit cluster assignments, and as such can help analysts in identifying “natural” group neighborhoods. This is the reason for showing the map view in connection and coordinated with the small multiples matrix view: inspecting both in combination provides the analyst with insights needed to decide whether the explicit cluster assignments should be modified and whether alternative cluster models should be investigated.

Several multidimensional projection techniques could be employed to obtain the similarity maps. Our choice of IDMAP is based on a comparative empirical analysis of data maps obtained with different techniques. We employed Neighborhood Preservation (NP) (PAULOVICH; MINGHIM, 2008) curves to compare techniques regarding their precision in preserving multidimensional neighborhoods. One such comparison is illustrated in Figure 39, on the ionospheric scintillation data. For varying neighborhood sizes k , the NP curve is computed as follows: the k -nearest neighbor sets of each data point in multidimensional and projected space are obtained, and their overlapping rate is computed, yielding values in the interval $[0, 1]$. The average precision values for all data points for the considered range of k are plotted. The resulting curves will approach maximum precision as k increases, however some techniques can achieve high precision even for smaller values of k such as the IDMAP projection.

Figure 39 – Neighborhood Preservation curves depicting precision of similarity maps of the 70 ionospheric scintillation variables obtained with four distinct projection techniques. Clearly, IDMAP yields precision rates superior to LSP and comparable to those obtained by classic approaches such as Fastmap and PCA.

Figure 39 shows the NP curves of similarity maps obtained with IDMAP and three alternative projection techniques, namely Fastmap, Least Squares Projection, and Principal Component Analysis (JOLLIFFE, 2002). Clearly, IDMAP yields precision rates superior to Least Square Projection (LSP) and comparable to those obtained by classic approaches such as Fastmap and PCA. Our experiments have indicated that IDMAP preserves neighborhood well for this data.

4.5.3.1 Hierarchical Similarity Map of Variables

In order to reduce the undesirable effect of circle overlapping, the map may be explored at multiple levels of detail. We propose a KD tree-based (SAMET, 1990) spatial representation of the 2D variable embedded points. The root contains projected variable points of all space, and the leaf nodes individual variables, as described in Figure 40. The map is shown initially at the lowest tree level, but the user can navigate up and down the tree hierarchy to reduce or increase the level of detail. At upper tree levels the map shows bigger circles that represent multiple variables. Colored circles representing multiple variables assigned to a single cluster are rendered with the cluster color, whereas circles rendered in gray represent a heterogeneous group of variables that have been assigned to multiple clusters. Hovering over a circle prompts the display of further details about the variables it represents.

Figure 40 – Schematic view of the hierarchical similarity map of variables. A KD tree spatial representation (SAMET, 1990) of the 2D embedded variable points. The root contains projected variable points of all space, and the leaf nodes individual variables

4.6.4 Variable Correlation Matrix

A correlation matrix of the aggregated variable's time matrices is computed, for all variables or for a user selected subset (Figure 41). Pairwise correlations are computed using Pearson correlation coefficient, as implied directly by the data values rather than the variable's feature vectors.

Figure 41 – Correlation matrix of Pearson coefficients computed from aggregated scintillation variables. Matrix rows and columns are sorted according to the clusters of variables, indicated by the colored lines along matrix borders.

Given the values of two aggregated variables $\mathbf{x}_i^{\text{tse}}$ and $\mathbf{x}_j^{\text{tse}}$, their Pearson correlation is given by Equation 4.24.

$$\rho(\mathbf{x}_i^{\text{tse}}, \mathbf{x}_j^{\text{tse}}) = \frac{\text{cov}(\mathbf{x}_i^{\text{tse}}, \mathbf{x}_j^{\text{tse}})}{\sigma_{\mathbf{x}_i^{\text{tse}}} \sigma_{\mathbf{x}_j^{\text{tse}}}}, \quad (4.24)$$

in which cov denotes the covariance and σ denotes the standard deviation.

Variable correlation is displayed as a heatmap visualization, where each cell depicts an entry of the correlation matrix. Values are mapped to a color scale where darker blue shades indicate strong negative correlations and darker brown shades indicate high positive correlations.

Matrix rows and columns are sorted based on the clustering of the variables, clusters are indicated by the colored borders at the top and the left side. As correlations are computed directly from the retrieved data observations, this visualization provides a complementary view to verify cluster cohesion. Analysts may inspect it to assess the similarity of variables assigned to a single cluster. These multiple visualizations focused on variable behavior allows analysts to investigate, assess and modify the cluster assignment of specific variables, to efficiently identify groups of variables with similar temporal behavior.

Normalized time series line graphs of the variables can also be inspected, with lines colored according to its variable's cluster. This view is associated with an array showing the corresponding matrix values, simulating a timeline, as shown in Figure 42, our version of the temporal pixel-oriented layout proposed by Machado *et al.* (2017). These views facilitate identifying common patterns between values in different time windows.

Figure 42 – Normalized time series graphs and timelines depicting the corresponding time-matrix values.

4.7 Visualization of Data Instances

Besides the visualizations focused in the variables, it is also relevant to display information on the behavior of the data instances. In particular, 2D similarity maps can convey groups of similar observations. Multiple maps may be created considering the data instances described by alternative variable spaces, so that users may compare the behavior of distinct descriptor spaces in relation to a target variable.

Several multidimensional projection techniques were considered to create the similarity maps of the data instances, and after exhaustive empirical investigation, we chose the Least Square Projection (LSP) technique for this purpose. LSP has been introduced by Paulovich *et al.* (2008a), and seeks to preserve local neighborhoods identified in the original data space. In particular, it is known to yield excellent precision in terms of neighborhood preservation at a reasonable polynomial computational cost (PAULOVICH *et al.*, 2008a). Initially, it projects a subset so-called “control points” with a dimension reduction method that preserves distance relationships accurately. Then, it constructs a linear system considering point neighborhoods in the original space and the projected coordinates of the control points. The solution of this linear system defines the remaining positions in the low-dimensional space.

In the resulting map each data instance is shown as a colored circle, where colors map either a category or a scalar variation of a selected target variable. Pairwise distances may be computed with the Euclidean distance or the Inverse Pearson correlation as a measure of

dissimilarity. Analysts can filter out data instances based on the value of the target variable, which is useful to reduce point overlapping and conduct more focused data analysis.

If data instances have an associated class label, the Silhouette Coefficient of the map may be computed considering the classes, which provides a quality measure of its class cohesion and separability. For example, it is helpful to determine which variable space is more effective in visually segregating data instances from different categories. These maps support a feature space analysis in order to identify the most representative variables for characterizing a behavior. As a default, a representative subset of variables is defined as the set of cluster medoids, however, this initial subset may be freely modified by the analyst after interacting with the multiple visualizations.

4.8 Chapter Notes

In this chapter the components of our *TV-MV Analytics* visual framework were described to support exploratory data analysis tasks on time-varying multivariate data on multiple temporal scales. Our framework offers multiple visualization and analytics resources for investigating the interplay of variables and/or data instances, and allows flexible navigation to maintain the global and detailed temporal context of each variable, in itself and in relation to others. The proposed visualizations of variables include (i) a hierarchical similarity map of their feature vectors, (ii) time matrices views for the variables, (iii) correlation matrices from aggregated variables, and (iv) time circular diagram. The variables can also be explored as aggregated and normalized line-graphs to convey seasonal patterns and hierarchical similarity-based visualizations of the user-defined data instances. All views are interactive in order to include the user-knowledge in the exploration process.

In the following chapter detailed case studies are presented reporting on the application of the framework to support analysis of data from four different domains, described in Chapter 3: ionospheric scintillation, crime statistics, air quality, and stock market. The ionospheric scintillation case studies were conducted in partnership with domain experts to provide supporting scenarios illustrating how *TV-MV Analytics* can be employed effectively to discover relevant information about the scintillation phenomenon.

EXPERIMENTAL RESULTS

5.1 Initial Remarks

Prototypical data analysis questions relative to the ionospheric scintillation (see Section 5.2) were defined in collaboration with domain experts, which provided the grounds for developing the analytics framework. The proposed solution was validated by means of case studies driven directly by these questions, using the functionalities and visualizations described in Chapter 4. In these case studies, we investigated the behavior of multiple variables recorded at a specific satellite and the same variables recorded at all satellites monitored from a specific station. Results confirmed that the *TV-MV Analytics* framework can be successfully applied to assist domain experts in answering relevant questions on the scintillation data.

Although *TV-MV Analytics* was designed with the ionospheric scintillation data in mind, the proposed analytics strategies can successfully be applicable to any other domains where large time-varying multivariate data arise and must be understood. We verified its potential on case studies on the data sets described in Chapter 3. For these, we considered some speculative questions and scenarios, and investigated alternative choices of feature vectors and similarity measures. In the case studies we used classifiers and/or regressors to assess alternative space representations and compare the behavior of multiple variables and assess their role in characterizing a variable of interest. We limit the discussions to selected motivating and illustrative examples, considering different exploration options and present the results obtained on the ionospheric scintillation data in Section 5.2, and on the other data domains in Section 5.3.

5.2 Visual Analytics of Ionospheric Scintillation Data

We departed from prototypical analysis questions, defined in partnership with expert collaborators, listed in Table 18, to define the five illustrative exploration scenarios described in this Section. For this investigation we downloaded into a local relational database two subsets of

the ionospheric scintillation data instances: observations from the Satellite GPS 1 recorded at the PRU1 station during January, 2014; observations from the 32 GPS satellites from the PRU2 station from January to December, 2015. January is interesting for investigation because it is when the summer season is at the peak in the southern hemisphere, and scintillation occurrence is typically high. We also investigate the behavior of the variable *s4_corrected* from the 32 GPS satellites of the PRU2 station. Euclidean distance is used to measure the (dis)similarity and the color moments are employed as time matrices feature vectors. The visualizations considered in this section are the time matrices and associated similarity map of variables, correlation matrix of variables (aggregated values), similarity map of instances and normalized time series (aggregated values).

Table 18 – High-level analysis questions of ionospheric scintillation data identified in collaboration with domain experts.

Q1	Which variables tracked show similar behavior over a target period?
Q2	How do variables relate to each other and to the scintillation indices?
Q3	At what level of detail do significant changes in the S4-value become evident?
Q4	Besides the S4, what (minimal) subset of variables suffices to characterize scintillation?
Q5	Can variable behavior be extrapolated in time?
Q6	Do the same variables tracked by different satellites show similar behavior?

5.2.1 Identifying Variables with Similar Temporal Behavior (Q1)

Data was obtained by sampling the maximum values recorded over one-hour periods along each day, from January 1-31, 00:00h-23:00h, with an elevation cutoff of 30 degrees, yielding 8,868 data instances described by 70 variables, without considering variables categorized as identifiers.

Clustering the variables with X-means and Euclidean distance produced seven clusters, with a Silhouette Coefficient $SC = 0.57$, which indicates a high-quality clustering result. The initial resulting model already reveals variable clusters with similar behavior over the selected period. We first inspect the combined views provided by the similarity map of variables and the multiple time matrices view, shown in Figure 43. While the map provides an overview of the variables' global similarities regarding their temporal behavior over the observed period, the small multiples view conveys their temporal behavior.

These views, combined with the correlation matrix, enable confirming (dis)similarities among cluster elements. For instance, in cluster C4 (lilac), we notice that *azim* behaves differently than *l5_e5b_locktime*, *l2_e5a_locktime*, *l1_locktime* and *f2nd_tec_locktime*, which present high positive correlation, which can be confirmed in the similarity map. The difference is also noticeable observing the lilac cluster in the correlation matrix view depicted in Figure 44.

Figure 44 – Variable Correlation Matrix (Pearson coefficient) for a specific subset of variables assigned to three distinct clusters (C4-lilac, C5-red and C6-brown).

The red cluster (C5) contains two subgroups with clearly distinct behaviors. The first subgroup contains all variables with prefix *avgccd_**, and the second one, the variables with prefix *avg_cn0** and *f2nd_tec_cn0*. In the similarity maps, the subgroups are separate, also evident when observing these variables in the correlation matrix view. Cluster C6 (brown) has two separate subgroups visible in the map, one of them formed by variables with prefix *p_l* and the other one with variables related to *s4* and the variable *numerator_si_l5_e5b*. In the time matrices, these variables do exhibit a different behavior, also observed in their correlation matrix.

These two visualizations when viewed in combination provide useful information for the analyst to adjust the cluster model obtained with X-means. The variables' cluster assignment can be modified using the combo box provided in its time matrix view. The ability to identify clusters of variables with similar behavior is important when investigating answers to the questions.

Considering the previous remarks, the cluster model was modified as follows: cluster C5 (red) was split into two subgroups formed by the two cohesive groups of three variables and four variables, respectively; and cluster C6 (brown) was also split into two groups. The adjusted cluster model, shown in Figure 45, is formed by 9 clusters and has an improved silhouette coefficient $SC = 0.61$. This modified clustering more appropriately represents a user's "vision" of the scenario, and can yield better classification results. Besides, the variable clusters in the adjusted model convey those variables with similar behavior over the selected period.

Figure 45 – IDMAP similarity map view of the same 70 variables depicted in Figure 43.A, in which circle colors reflect an alternative (user-adjusted) cluster model (9 clusters).

5.2.2 Identifying Relations between S4 Indices and Other Variables (Q2) and Relevant Variables to Scintillation (Q4)

Departing from the previous study and in order to investigate how the multiple tracked variables relate with the S4 indices, we compared the results of classifiers on the set of data instances represented with three alternative variable spaces, taking *S4_corrected* as target variable and the Tiwari categorization (see Section 3.2) as ground truth, because it is one of the most used and recommended by our collaborator specialists.

The three spaces considered as descriptors are formed by (a) all 58 monitored variables that are not directly derived or theoretically associated with measures of the S4 index (obtained discarding the 12 variables with prefix *s4* and *vs4*); (b) the representative variables defined by the cluster medoids resulting from X-means (7 variables) clustering, also depicted in Figure 43; and (c) the representative variables (medoids) defined by the user-adjusted cluster model, shown in Figure 45, obtained after observing and interacting with the visualizations. The X-means clustering of all 70 variables (Figure 43) generated seven clusters, which have as their corresponding medoids the variables: *sigmaccd_l1*, *dtec_4530*, *phi10_l5_e5b*, *l1_locktime*, *avg_cn0_l5_e5b*, *si_l2c_e5a* and *numerator_si_l5_e5b*. The user-adjusted clustering (Figure 45) has nine clusters with medoids: *sigmaccd_l1*, *dtec_4530*, *phi10_l5_e5b*, *l1_locktime*, *avg_cn0_l5_e5b*, *si_l2c_e5a*, *numerator_si_l5_e5b*, *avg_cn0_l1* and *p_l1*.

A similarity map of the data instances described by 58 variables (all variables except measures of the S4 index) is shown in Figure 46a, which has $SC = 0.35$, in which circle colors map the values of index *S4_corrected* with the heated-object color map. Despite the severe overlapping, one notices that data instances associated with higher scintillation appear towards the right region.

The resulting map of observations given by the descriptor space obtained from the X-Means clustering is shown in Figure 46b, $SC = 0.35$. Finally, the resulting map of the observations given by the user-adjusted descriptor is shown in Figure 46c, has a slightly improved silhouette $SC = 0.38$. These results suggest the reduced spaces, given by seven or nine variables, can be sufficient to characterize scintillation behavior.

Figure 46 – Similarity maps of 8,868 data instances described by (a) all variables except measures of the S4 index (total of 58 variables), (b) seven variables defined by the cluster medoids, and (c) nine user-defined variables defined after interaction. Circle colors map scintillation intensity as given by index $S4_{corrected}$, with higher intensities indicated by darker colors.

We further investigated this hypothesis, i.e., the relevance of these two reduced subsets of variables to characterize scintillation behavior in comparison to the 58 variables, by running classifiers and a linear regression on the set of data instances described by the three spaces, taking $S4_{corrected}$ as target variable and the Tiwari categorization as ground truth. We also applied a linear regression to the different variable subsets, to compare the predicted and the actual values.

Results (classifier accuracy and regression correlation) are summarized in Table 19, confirming that the user selected subset of variables is indeed appropriate to characterize scintillation behavior. According to the scintillation experts, this is a relevant result, as most of these attributes could be monitored using conventional geodesic receivers available, for example, from the Brazilian Network for Continuous Monitoring of the GNSS Systems (RBMC), rather than the highly specialized and more costly ISMRs.

Table 19 – Classification accuracy and regression correlation of data instances described by three distinct variable subspaces.

Technique	58 variables	7 medoids	9 variables
J48	96.53%	97.42%	97.72%
Multilayer Perceptron	96.67%	97.58%	97.93%
Linear Regression	0.87	0.73	0.78

The behavior of such variables in comparison to $S4$ indices can be further investigated by means of normalized time series views of the aggregated data represented in the time matrices. For example, Figure 47 shows a sample of the normalized time series for variables avg_cn0_l1 and $S4_corrected$ over three days (from January 1 to 3) and the timeline considering time-matrix values, where daily patterns are observable for both variables.

Figure 47 – A three-day sample of the normalized time series of aggregated values (maximum over 1h intervals) for variables avg_cn0_l1 and $S4_corrected$, where a daily seasonal pattern can be observed.

5.2.3 Analysis at Multiple Temporal Granularities (Q3)

The granularity option allows varying the aggregation period over which the analysis is conducted. In order to illustrate this capability, we analyzed the same subset of data instances considered in the first case study (Section 5.2.1), now considering the maximum value observed time intervals of varying-length: three hours, two hours, one hour, 30 minutes, 20 minutes, 10 minutes and five minutes. Figure 48 summarizes the cluster models obtained, in terms of their size and composition (indicated by colors). One notices that cluster patterns differ over the different levels of granularity, although a few clusters remain quite stable, e.g., the ones shown in purple and light blue are quite well-preserved for all time intervals. The capability of investigating global relationships between variables considering different temporal granularities may assist experts in further investigating the phenomenon over longer time ranges.

Figure 48 – Cluster models of the 70 variables in terms of their size and composition (indicated by colors), obtained from data instances aggregated at different time intervals: three hours, two hours, one hour, 30 minutes, 20 minutes, 10 minutes and five minutes. Cluster patterns differ over the different levels of granularity, although a few clusters remain quite stable.

5.2.4 Variable Prediction (Q5)

In order to find out what extent the time matrix representation of variables can assist in prediction tasks, we conducted some experiments to verify our system's capability of predicting next-day values of variables from the information recorded in their time matrix representation. We employed isotonic regression to predict values of January 31, 2014, based on information between January 1-30, 2014. Results for variables *S4_corrected* and *Si_II* are shown in Figure 49, comparing the Pearson correlation between known and predicted values (hovering over the cells will display the actual values). Prediction of *S4_corrected* shows a correlation of 0.9731 with actual values, whereas for *Si_II* correlation is 0.8863, which are good results and confirm that these variables can be extrapolated in time.

Figure 49 – Actual and predicted values (January 31, 00:00h-10:00h) of variables (a) *S4_corrected*; (b) *Si_II*. Prediction of *S4_corrected* shows a correlation of 0.9731 with actual values, whereas for *Si_II* correlation is 0.8863.

5.2.5 Comparing Data Instances from Multiple Satellites (Q6)

Identifying groups of satellites more susceptible to scintillation over a region is particularly relevant to derive specific strategies for dealing with this effect on GNSS positioning. A typical task would be concerned with inspecting scintillation values recorded from multiple GPS satellites, and observe their consistency. This analysis may consider, for example, the *S4_corrected* index, which measures amplitude scintillation computed without ambient noise effects.

We considered data instances of the *S4_corrected* index recorded at the PRU2 station from January to December 2015, which were aggregated by taking the average daily value. This query returned 1,775,039 data instances from the 32 GPS satellites, which were SC optimized-K-means clustered picking $k = 8$ (which yielded the best silhouette, $SC = 0.498$). The resulting clustering (C1 to C8 identified by colors) is shown in the small multiples time matrix view in Figure 50.

Figure 50 – (A) IDMAP similarity map and (B) Small Multiple Time Matrices view of values of index *S4_corrected* recorded at the 32 GPS satellites (January 1 - December 31, 2015). Matrix cells show average daily values along each month. Numeric labels in the similarity map identify the satellites.

These clusters reveal distinct patterns of amplitude scintillation, which is typically more severe from November to March. This is confirmed by the time matrices, where the initial and/or final rows (corresponding to the first and last months of the year) show higher averages (darker

colors). The three satellites in the white cluster were seemingly less affected by scintillation over this period. The affected satellites at a particular GPS station will change according to the spatial dynamics of both their fixed orbital periods and the ionospheric irregularities. Time matrix views of four satellites (4, 8, 10, 26) depicting over 28% missing data are highlighted (cyan borders). Occurrence of missing data may be related to scintillation susceptibility of a certain satellite because strong scintillation may cause loss of lock and a consequent inability to track some variables. If this is the case, data is missed over short periods (typically a few minutes). One notices that these four satellites display extensive periods of missing data, more likely a consequence of satellite unavailability, e.g., due to maintenance or replacement operations in the GPS constellation.

5.3 Exploring Data from Other Domains

We present results obtained using our framework to investigate the other time-varying multivariate data sets described in Chapter 3: we conducted four exploratory studies on the crime statistics of São Paulo state, Brazil, one study on the air quality data, and one study on the stock market data.

5.3.1 Crime Statistics of São Paulo State, Brazil

We considered five illustrative data analysis questions, listed in Table 20, developed in four exploratory case studies, to investigate this data set. For this purpose we created a local relational database with the monthly crime statistics at the fourteen regions of São Paulo from January 2002 to February 2017. Dissimilarity was measured with the Inverse Pearson Correlation and the Abo-Zaid moments employed as the time matrices features, which emphasize their variations.

Table 20 – Analysis questions of crime statistics of São Paulo State, Brazil.

Q1	How do indicators of particular types of crime behave on different regions?
Q2	Given a target indicator, how does its behavior compare over the different regions?
Q3	Did fire armed police indeed impact crime rates in São Paulo? (ARVATE; SOUZA, 2011)
Q4	How do crime indicators relate to homicide rates?
Q5	Can the behavior of crime indicators be extrapolated in time?

5.3.1.1 Comparing Crime Indicators of Three Regions (Q1)

We explored three indicators related to drug occurrences (drug possession, drug seizure and drug traffic) and the indicator illegal weapon possession recorded at three regions: São Paulo capital, São Paulo metropolitan area, and Ribeirão Preto, a city with strong economical activity

Figure 51 – Occurrences of drug possession, drug seizure, drug traffic and illegal weapon possession in three regions (from January 2002 to February 2017), São Paulo capital, São Paulo metropolitan area, and Ribeirão Preto: (A) IDMAP similarity map view; and (B) small multiple time matrices view of the 12 indicators.

within the state of São Paulo. The monthly data over the period yields 181 data instances, and 12 indicators are observed in total (four associated with each of the three regions).

We considered different algorithms to obtain an initial clustering of the indicators to start exploration, the respective silhouette coefficients are shown in Table 21. The K-means algorithm with $k = 3$ yielded the best model, with $SC = 0.61$. The combined views of the similarity map of indicators and the multiple time matrices are shown in Figure 51.

Table 21 – Silhouette coefficient(SC) of different clustering models of four indicators related to occurrences of drug (3) and illegal weapon possession in the crime statistics data.

Algorithm	# Clusters	SC
X-means	2	0.37
K-means	2	0.45
	3	0.61
	4	0.28
	5	0.05
B-kmeans	2	0.47
	3	0.48
	4	0.43
	5	0.38
K-medoids	2	0.47
	3	0.26
	4	0.50
	5	0.05

Figure 52 – IDMAP similarity map view of the same indicators depicted in Figure 51.A, in which circle colors reflect an alternative (user-adjusted) cluster model (4 clusters).

They show that occurrences of drug possession present a similar behavior at the Capital and the Metropolitan area regions (blue cluster C1), but a different behavior in Ribeirão Preto (the salmon cluster C2). Cluster C2 also includes the indicators related with drug-traffic at the three regions, drug seizure at the Capital and at Ribeirão Preto. The indicator of illegal weapon possession behaves similarly in the three regions, and the three were assigned to cluster C3 (purple). The drug-traffic indicators at the three regions assigned to cluster C2 do exhibit a distinct behavior. Considering this remark, the cluster model was modified as follows: a new cluster was included grouping only the drug-traffic variables at the three regions, the remaining three clusters being adjusted accordingly. The modified cluster model, illustrated in Figure 52, is

formed by 4 clusters with $SC = 0.49$. Although its silhouette coefficient is not improved, this model could better represent a user's "vision".

5.3.1.2 Same Variable from Different Regions (Q2) and Variable Prediction (Q5)

Figure 53 – Time Matrices of *total number of inquiries* by region associated to the Map of São Paulo, Brazil.

We also investigated the variable *total number of inquiries* for the 12 regional subdivisions of the São Paulo state, shown in Figure 53, again using the Inverse Pearson Correlation to measure dissimilarity and the Abo-Zaid moments in order to verify which spatial regions present similar behavior on different indicators. Clustering, obtained with the SC optimized-K-means result, produced five clusters ($SC = 0.6$): the blue cluster (C1) with the Capital and metropolitan São Paulo (two spatially adjacent regions); the beige cluster (C2) includes regions Sorocaba and Araçatuba; the brown cluster (C3) Campinas, Riberão Preto and São José do Rio Preto (of which the later two spatially adjacent); the light green cluster (C4) includes Presidente Prudente, Bauru and Piracicaba; and finally regions Santos and São José dos Campos are in the purple cluster (C5). We colored the regions in the map of São Paulo shown in the Figure according to their assigned clusters.

Figure 54 – Map of the state of São Paulo conveying the clustering model for each crime indicator.

We also investigated the spatial behavior of other indicators, as shown in Figure 54, using the same settings: Inverse Pearson Correlation, the Abo-Zaid moments and the Silhouette-K-means. In most clustering results, indicators from São Paulo capital and São Paulo metropolitan area belong to the same cluster, except for five indicators: (c) Number of arrests offenders by warrant, (e) Number of detentions of offenders by warrant, (i) Seizure of narcotics, (k) Seized firearms and (l) Criminal possession of weapons. The clustering spatial patterns are different in each case, suggesting that indicators present distinct global behaviors in each region.

We considered data from January 31, 2002 to December 31, 2015, to predict values of indicator *imprisonments* for each region in 2016; correlations between actual and predicted values are listed in Table 22. Figure 55 summarizes the actual and the predicted values for the state as a whole, with a strong correlation of 0.76.

Table 22 – Correlations between actual and predicted values of indicator *imprisonments* using historical data from January 31, 2002 to December 31, 2015. Most correlations are positive, which indicates that it is possible to predict the values of this indicator. However, the Santos indicator has a negative correlation, which would indicate the presence of other factors.

Region	Correlation of Prediction
São Paulo capital	0.51
São Paulo metropolitan area	0.52
Araçatuba	0.27
Bauru	0.75
Campinas	0.65
Piracicaba	0.63
Presidente Prudente	0.65
Ribeirão Preto	0.31
Santos	-0.19
São José do Rio Preto	0.83
São José dos Campos	0.56
Sorocaba	0.53
Total State	0.76

Figure 55 – Actual and predicted values of *imprisonments* at the state of São Paulo in 2016. The correlation between actual and predicted is 0.76.

5.3.1.3 Impact of Fire Armed Municipal Police (Q3) and Relationship between Crime Indicators and Homicide Rates (Q4)

We investigated the six indicators considered in the study by [Arvate and Souza \(2011\)](#): total number of inquiries, number of arrests of flagrant offenders, seizure of narcotics, number of recovered vehicles, number of imprisonments and homicide rates. They considered data from 2002 to 2011 to estimate the impact of fire-armed police on crime in Brazil, after legislators approved a law to regulate the use of firearm by municipal polices. They concluded that fire armed municipal police contributed to reduce crime indicators significantly after 2003.

Figure 56 – Time Matrices of the six indicators: Total number of inquiries, number of arrests of flagrant offenders, seizure of narcotics, number of recovered vehicles, number of imprisonments and homicide rates (from January 31, 2012 to February 28, 2017).

Our exploratory investigation covers a longer period, from January 31, 2012 to February 28, 2017 (time matrices are shown in Figure 56), in order to verify whether their conclusion holds. Three clusters were identified ($SC = 0.44$) using the same settings: Inverse Pearson Correlation, the Abo-Zaid moments employed as the time matrices features and Silhouette-K-means clustering. A blue cluster (C1) includes the indicators on *number of arrests of flagrant offenders* and *seizure of narcotics*, a yellow cluster (C2) includes *number of imprisonments* and *homicide rates*, and a pink cluster (C3) includes *number of recovered vehicles* and *total number of inquiries*. We observe a reduction in all the indicators along a small period after 2003, but a significant increase is observed from 2011, except for the indicators *homicide rates* and *number of recovered vehicles*, which display different behaviors. The *number of recovered vehicles* decreased in 2007 and 2008, but increased again later; and the *homicide rates* decreased after 2004.

As clusters were computed using the inverse of Pearson Correlation as dissimilarity measure, we found that cluster C2 actually conveys the negative correlation of both indicators *number of imprisonments* and *homicide rates*, i.e., the number of reported *homicide rates* decreased as the *number of imprisonments* increased (this is noticeable in the patterns of both time matrices in the yellow cluster). This suggests the alleged positive impact of fire armed municipal police might have been of short duration, and other factors may be causing an increase in the crime indicators reported. Such a reasoning is also supported by the time series graphs depicting the behavior of these indicators in Figure 57, in which reported numbers increase after 2010.

Figure 57 – Time series graphs of the six indicators: *number of arrests of flagrant offenders*, *seizure of narcotics*, *number of imprisonments*, *homicide rates*, *number of recovered vehicles* and *total number of inquiries* recorded from January 31, 2012 to February 28, 2017. For most indicators there is an increase after 2010.

We also investigated the role of the remaining indicators for characterizing the behavior of the *homicide* indicator. We initially created the small multiples time matrices view combined with the IDMAP similarity map for the 14 indicators over the São Paulo State (Figure 58). Data was obtained by sampling the values recorded over one-month periods along each year from January 2002 to February 2017.

Clustering the variables with SC optimized-K-means and the Inverse Pearson Correlation produced four clusters, with $SC = 0.51$, which indicates a high-quality clustering. The views enable confirming the correlation among cluster members. For instance, in cluster C1 (black) we notice that the three indicators *official reports of flagrant crimes*, *imprisonments* and *number of arrests of flagrant offenders* are highly correlated. The indicators in clusters C2 (rose) and C3 (orange) present weak correlation, which can be confirmed in the similarity map. The cluster C4 (yellow) joins two groups with a strong positive correlation, a first group containing *criminal possession of weapons*, *seized firearms* and *homicide rates*, and a second group containing *seizure of narcotics* and *traffic of narcotics*, and these two groups present high negative correlation. The selected medoids were: *official reports of flagrant crimes*, *number of detentions of offenders by warrant*, *number of arrests of offenders by warrant* and *seized firearms*. Considering this remark, cluster C4 was split into two subgroups. The adjusted model, shown in Figure 59, is formed by 5 clusters and has an improved $SC = 0.56$. A new medoid was added to the representative indicators: *seizure of narcotics*.

Visually, the indicators most correlated with *homicide rates* are *criminal possession of weapons* and *seized firearms*. Departing from this analysis, in order to further investigate this issue, we compared the results of linear regressors on the set of data instances described by alternative spaces of indicators, with indicator *homicide* as target. The following spaces were considered as descriptors:

1. **All indicators:** All 13 indicators, excluding *homicide rates*.

Figure 58 – 14 indicators for the São Paulo State: (A) IDMAP similarity map view, with each circle depicting a variable; and (B) small multiple time matrices view, where an entry in each matrix shows the maximum value recorded over a monthly period along each year.

Figure 59 – IDMAP similarity map view of the same 14 indicators depicted in Figure 58, now considering a user-adjusted cluster model (5 clusters).

2. **Medoids (4) of the initial automatic cluster model:** *official reports of flagrant crimes, number of detentions of offenders by warrant, number of arrests of offenders by warrant and seized firearms.*
3. **Medoids (5) of user-adjusted cluster model:** *official reports of flagrant crimes, number of detentions of offenders by warrant, number of arrests of offenders by warrant, seized firearms and seizure of narcotics.*
4. **Indicators obtained from the automatic linear regression (9):** *traffic of narcotics, criminal possession of weapons, seized firearms, official reports of flagrant crimes, number of detentions of flagrant offenders, number of arrests of flagrant offenders, number of arrests of offenders by warrant, imprisonments and recovered vehicles.*
5. **Indicator obtained from the automatic simple linear regression:** *criminal possession of weapons.*
6. **Indicators considered by Arvate and Souza (2011) (5):** *number of arrests of flagrant offenders, seizure of narcotics, number of imprisonments, number of recovered vehicles and total number of inquiries.*
7. **Indicators obtained from the automatic linear regression of the variables considered by Arvate and Souza (2011) (3):** *seizure of narcotics, number of imprisonments and number of recovered vehicles.*
8. **Indicators obtained from the automatic simple linear regression of the variables considered by Arvate and Souza (2011):** *seizure of narcotics.*

Results are summarized in Tables 23 and 24, in which the figures suggest that multiple variable spaces can properly characterize the indicator on homicide rates. Notice in most similarity maps high homicide rates (represented by dark circles) are placed distant from the low homicide rates (light colored circles), confirming that the visual interaction in fact assist in identifying the best subset of indicators to characterize the *homicide* indicator.

Table 23 – Regression correlation of data instances described by the distinct indicator subspaces. # is the space number, # ind. the number of selected indicators and corr. the correlation.

#	# ind.	indicators	corr.	Projection
1	13	<i>possession of narcotics</i> <i>traffic of narcotics</i> <i>seizure of narcotics</i> <i>possession of weapons</i> <i>seized firearms</i> <i>official reports</i> <i>detentions of flagrant</i> <i>detentions by warrant</i> <i>arrests of flagrant</i> <i>arrests by warrant</i> <i>imprisonments</i> <i>recovered vehicles</i> <i>total inquiries</i>	0.97
2	4	<i>official reports</i> <i># detentions by warrant</i> <i># arrests by warrant</i> <i>seized firearms</i>	0.94
3	5	<i>official reports</i> <i># detentions by warrant</i> <i># arrests by warrant</i> <i>seized firearms</i> <i>seizure of narcotics</i>	0.94
4	9	<i>traffic of narcotics</i> <i>possession of weapons</i> <i>seized firearms</i> <i>official reports</i> <i># detentions of flagrant</i> <i># arrests of flagrant</i> <i># arrests by warrant</i> <i>imprisonments</i> <i>recovered vehicles</i>	0.97

Table 24 – Regression correlation of data instances described by the distinct indicator subspaces. # is the space number, # ind. the number of selected indicators and corr. the correlation.

#	#ind.	indicators	corr.	Projection
5	1	<i>possession of weapons</i>	0.95
6	5	<i># arrests of flagrant seizure of narcotics imprisonments recovered vehicles total number of inquiries</i>	0.82
7	3	<i>seizure of narcotics imprisonments recovered vehicles</i>	0.83
8	1	<i>seizure of narcotics</i>	0.72

5.3.2 Air Quality Data of an Italian City

In this data set, which has a small number of variables, we investigated the multiple variables to identify which ones present similar behavior over time, considering the Inverse Pearson Correlation to measure the (dis)similarity. We noticed the temporal matrices presented constant small variations, unlike the previous data sets explored. Thus, we adopted as features the orthogonal variant moments (OVM), which are sensitive to perturbations.

Bearing in mind that data instances describe hourly average values for a total of 13 variables measured by two commercial sensors (to measure temperature and absolute and relative humidity), the Metal Oxides (MOX) sensing device (composed by five sensors measuring MOX_{CO} , MOX_{NMHC} , MOX_{NO_x} , MOX_{NO_2} and MOX_{O_3}) and the conventional sensing device (spectrometer analyzers) which provides reference data for some pollutants: ST_{CO} , ST_{NMHC} , $ST_{C_6H_6}$, ST_{NO_x} and ST_{NO_2}).

Four clusters were found ($SC = 0.8$) with the SC optimized-K-means algorithm, where variables in each cluster present a visible correlation. The Blue cluster (C1) includes the variables absolute humidity, ST_{NO_2} and temperature. Absolute humidity and temperature variables show a positive correlation, and both are negatively correlated to ST_{NO_2} . The Red cluster (C2) includes $ST_{C_6H_6}$, ST_{NO_x} and MOX_{NO_x} with similar behavior. ST_{CO} and MOX_{NMHC} form the lavender cluster (C3). The beige cluster (C4) includes polluting gases measurements from the MOX sensor MOX_{CO} , MOX_{NO_2} and MOX_{O_3} related to variable relative humidity.

Figure 60 – Time Matrices of air quality data considering the entire period March 2004 - April 2005.

We used the simple linear regression function to identify the most correlated variable with each of the five polluting gases. The results, listed in Table 25, show that the measurements for most polluting gases are highly correlated, which is also confirmed in the time matrices view.

Vito *et al.* (2008) found high correlation between some variables (see Table 26 column A), considering the hourly samples for all period (in Table 26 they just considered measurements from the initial eight days, as the ST_{NMHC} analyzer went out of service). We also explored the correlation with simple linear regressions but considering daily aggregations of the same

Table 25 – Most correlated variables with each polluting gas.

Polluting gas	Variable	Correlation	Equation
$ST_C_6H_6$	MOX_NMHC	0.98	$ST_C_6H_6 = 0.03 * MOX_NMHC - 15.66$
ST_CO	ST_NO_x	0.79	$ST_CO = 0.01 * ST_NO_x + 0.83$
MOX_CO	MOX_O_3	0.9	$MOX_CO = 0.49 * MOX_O_3 + 598.69$
ST_NO_2	ST_NO_x	0.76	$ST_NO_2 = 0.17 * ST_NO_x + 70.3$
MOX_NO_2	MOX_NMHC	0.78	$MOX_NO_2 = 1.01 * MOX_NMHC + 509.16$
ST_NO_x	ST_NO_2	0.76	$ST_NO_x = 3.36 * ST_NO_2 - 133.11$
MOX_NO_x	ST_NO_x	0.66	$MOX_NO_x = -0.82 * ST_NO_x + 1022.27$
MOX_O_3	MOX_CO	0.9	$MOX_O_3 = 1.67 * MOX_CO - 809.54$

data, both to confirm the correlation coefficients and also to define the best regression equations to predict the values (Table 26, column B).

Table 26 – Correlations between measurements of different pollutants obtained from the reference device (spectrometer analyzer) as computed by Vito *et al.* (2008) and as obtained in our framework. Our results, obtained through interacting with the views, confirm that aggregated data instances suffice to characterize the target variables.

		A	B	
variable 1	variable 2	Vito	our result	equation
ST_NMHC	$ST_C_6H_6$	0.98	0.98	$ST_NMHC = 35.17 * ST_C_6H_6 + 584.52$
ST_CO	ST_NO_x	0.78	0.79	$ST_CO = 0.01 * ST_NO_x + 0.83$
ST_CO	ST_NO_2	0.67	0.68	$ST_CO = 0.02 * ST_NO_2 - 0.17$
$ST_C_6H_6$	ST_NO_x	0.72	0.72	$ST_C_6H_6 = 0.03 * ST_NO_x + 4.09$
$ST_C_6H_6$	ST_NO_2	0.60	0.61	$ST_C_6H_6 = 0.1 * ST_NO_2 - 0.4$
ST_NO_x	ST_NO_2	0.76	0.76	$ST_NO_x = 3.36 * ST_NO_2 - 133.11$
ST_CO	$ST_C_6H_6$	0.90	0.93	$ST_CO = 0.18 * ST_C_6H_6 + 0.28$

TV-MV Analytics allows further investigating a variable correlation matrix visualization. We inspected the correlation matrices of the original data (hourly values), see Figure 61.a; and of the aggregated data (daily maximum), see figure 61.b, to assess the impact of the data reduction due to aggregation. After aggregation, the time series have been reduced from 9,384 to 391 samples. In Figure 62 it is possible to observe the higher correlation between $ST_C_6H_6$ and MOX_CO over a short period (Figure 62.a) and a longer period for the aggregated data (Figure 62.b). Also, some patterns can be found in Figure 62.a in the days June 18-21 for MOX_CO and in the days June 17-21 for C_6H_6 . In Figure 61, we observe that in spite of correlation values are slightly different, the global patterns are preserved.

The correlations between the measurements obtained from the spectrum analyzers (the reference device) and the multisensor device (MOX), for each pollutant were computed. Notice that values are not correlated, which can be confirmed observing the time matrices shown in Figure 60. ST_NO_2 and MOX_NO_2 have a correlation of 0.16; ST_NO_x and MOX_NO_x have a

correlation of 0.66 and ST_CO and MOX_CO have a correlation of 0.89.

Figure 61 – Correlation Matrix of original and aggregated values measured for variables: C_6H_6 , plus CO , NO_2 and NO_x (reference and MOX), from March 10, 2004 to April 04, 2005. Values are mapped to a color scale where darker blue shades indicate strong negative correlations and darker brown shades indicate high positive correlations.

Figure 62 – Line graphs depicting the normalized series of measurements of $ST_C_6H_6$ (pink) and MOX_CO (blue): (a) original hourly values (May 15 to May 22, 2004) and (b) aggregated daily values (April to August, 2004).

We also explore the correlations between values over sub-periods of one day with the time circular diagram for variables $ST_C_6H_6$ and MOX_CO in Figure 63, confirming patterns found in Figure 62 in the days May 18-21.

Figure 63 – Time circular diagram of $ST_C_6H_6$ and MOX_CO , considering values from May 15 to May 22, 2004 and 1-day sub-periods.

5.3.3 US Stock Market Data

Table 27 lists the two data analysis questions investigated in the stock market domain. We used the OVM moments to represent the time matrices to highlight small variations (see Section 4.5.2), the inverse of Pearson Correlation as dissimilarity measure and considered the average monthly values from January 1, 1995 to November 30, 2016. For the analysis we limit the discussions to two selected subsets of information relative to the segment of technological companies: 12 leaders in the sector and the total 438 technological companies. The 12 companies from United States were considered the biggest technological companies in 2016 according to Forbes (SHARF, 2016) and are listed in the Nasdaq Stock Market, see Table 28.

Table 27 – Analysis questions of stock market data.

Q1	Which companies present similar behavior of <i>close</i> value after 2000?
Q2	Can we identify groups of companies with similar behavior? What characterizes these different behaviors?

Figure 64 shows the time matrices view of the variable *close* for the 12 companies. Six clusters were obtained with $SC = 0.82$. The Blue cluster (C1) includes AMZ and QCOM, where the initial series of values are similar, but the final ones differ considerably. The behavior of variable *close* of Apple (AAPL) is very unlike any other, and it has been assigned alone to the pink cluster (C2). The orange cluster (C3) includes CSCO and MSFT, which have very similar behavior. FB and IBM are in the dark blue cluster (C4), in spite of data from FB started being registered from 2012, the behaviors over the final period are similar. Google and HPQ in the red

Table 28 – US biggest technological companies in 2016 (according to Forbes) listed on Nasdaq Stock Market.

#	Company	Stock symbol
1	Apple	AAPL
2	Microsoft	MSFT
3	Alphabet/Google	GOOGL
4	IBM	IBM
5	Intel	INTC
6	Cisco Systems	CSCO
7	Oracle	ORCL
8	Hewlett Packard Enterprise	HPQ
9	Qualcomm	QCOM
10	Facebook	FB
11	Amazon	AMZ
12	Texas instruments	TXN

cluster (C5) show a negative correlation. And in the light green cluster (C6), the behavior of *close* for companies INTC, ORCL and TXN seems rather uniform.

In 2001, the US stock prices took a sharp downturn. All time matrices of the technological companies (except for Microsoft, which shows a delayed response), conveys this fact. After the September 11 attacks, the indices slid steadily from March 2002, with some dramatic declines occurring between July and September.

We employ the prediction functionality to predict the behavior of variable *close* on a particular year, considering different intervals of known historical data. For this purpose, we considered the variable *close* for IBM, whose time matrix is shown in Figure 65. The more information the model has, the more accurate the prediction will be (NORIEGA *et al.*, 2013). This is confirmed in the prediction results for variable *close* for IBM in 2016, considering different periods of historical data: 31, 15 and 5 years, shown in Table 29, where best results were obtained with the longer interval.

We explore the data relative to all 438 technological companies to identify the most representative ones along the tracked period (Q2), using the hierarchical version of the small multiple time matrices view of the variables, the color moments and Euclidean distance, highlighting the impact of missing data. Figure 66 shows the five clusters obtained with the K-means algorithm with $SC = 0.33$. Blue cluster (C1) contains 99 companies, lilac cluster (C2) 60 companies, red cluster (C3) 45 companies, olive cluster (C4) 121 companies and green cluster (C5) 113 companies.

Five groups of companies with similar behavior were identified, from which five medoids are identified as representative companies of each group pattern: TeleTech (TTEC), AOL, JDS Uniphase Corporation (JDSU), Rovi Corporation (ROVI) and Fortinet (FTNT). We also show the first two levels of the hierarchy to explore sub-representative companies, results are summarized

Figure 64 – Time Matrices and similarity map for the 12 US companies.

Table 29 – Prediction results of variable *close* for IBM in 2016, using different intervals of historical data: 31, 15 and 5 years.

historical data	correlation	result
31 years	0.51	<div><div>Reality</div><div>2016</div><div>Prediction - isotonicRegression</div><div>2016</div><div>Correlation: 0.5076</div></div>
15 years	0.46	<div><div>Reality</div><div>2016</div><div>Prediction - isotonicRegression</div><div>2016</div><div>Correlation: 0.4591</div></div>
5 years	0.13	<div><div>Reality</div><div>2016</div><div>Prediction - isotonicRegression</div><div>2016</div><div>Correlation: 0.1321</div></div>

Figure 65 – Historical data of variable *close* for IBM from January 1985 to December 2015. Three intervals of historical data were considered to predict the values for 2016.

Figure 66 – Hierarchical time matrices of the variable *close* for 438 companies (January 1, 1995 to November, 30 2016): (A) IDMAP similarity map view, with cluster medoids highlighted in bright green; and (B) small multiple time matrices view of the variables which are cluster medoids. Each matrix entry shows the maximum monthly value recorded along each year.

in Figure 67. Detailed views generated for nodes highlighted in red are shown in Figures 68, 69, 70, 71 and 72. Visually, the groups are well defined, where the companies show similar temporal behavior. In Figure 68, the red group, represented by the time matrix of company PCTI, shows a particular behavior, where high values are placed at the very beginning of the period (first two or three rows) and low values appear later, this behavior appears in all cluster members, which are: AMCC, BCOR, EGAN, EXTR, FNSR, INAP, IXYS, OCLR, PSEM, QCOM, RMBS, RNWK, SIMG, SONS, TQNT and TSYS.

The blue cluster includes the companies CNQR, FEIC, HSTM, ININ, NTCT, OSIS, ULTI and URI, and is represented by PCYG. In this group low values of *close* occur at the beginning and high values occur more towards the end of the observed period.

The yellow group is represented by the BRKS company and includes the companies ADI, AMD, AVNW, BBSI, CAMP, CREE, CTS, DDD, DELL, EBIX, EMC, ISSI, LTXE, MCRL, MDRX, MU, PLCM, PLXS, QLGC, ROVI, SEAC, SMTC, SWKS, VRSM and ZIGO. This cluster is characterized by a soft variation of intensity.

The RNG company is the representative of the lilac cluster, which includes the companies BRDR, CUDA, DWRE, EOPN, EVTC, GWRE, HPE, LNKD, NOW, NTK, PAYC, QTWO, RNG, RUBI, SPSC, SWI, TRIP and VRSK. Their behavior is characterized by a high percentage of missing data and peak values.

The green cluster is also characterized by a high percentage of missing data, but values are not as high, it includes the companies AYR, CAVN, CTCT, CVLT, GSIT, GUID, IMMR, MNTX, PERI, PRO, RAX, RBCN, TDC, TRAK, TSRA and VOCS, where PRO is the representative medoid.

Figure 67 – First and second hierarchy levels of the hierarchical time matrix structure depicting the US technological companies. Detailed views are provided for a subset of selected nodes (highlighted in red) (Figures 68 to 72).

Figure 68 – The red group: high values occur mostly at the beginning of the observed period, followed by low values. Companies: PCTI (medoid), AMCC, BCOR, EGAN, EXTR, FNSR, INAP, IXYS, OCLR, PSEM, QCOM, RMBS, RNWK, SIMG, SONS, TQNT and TSYS.

Figure 69 – The blue group: low values occur at the beginning and high values occur at the end of the period. Companies CNQR, FEIC, HSTM, ININ, NTCT, OSIS, ULTI and URI and PCYG (medoid).

Figure 70 – The yellow group: characterized by a smooth variation in intensities. Companies: BRKS (medoid), ADI, AMD, AVNW, BBSI, CAMP, CREE, CTS, DDD, DELL, EBIX, EMC, ISSI, LTXE, MCRL, MDRX, MU, PLCM, PLXS, QLGC, ROVI, SEAC, SMTC, SWKS, VRSM and ZIGO.

Figure 71 – The lilac group: a high percentage of missing data and high values. Companies: RNG (medoid), BRDR, CUDA, DWRE, EOPN, EVTC, GWRE, HPE, LNKD, NOW, NTK, PAYC, QTWO, RNG, RUBI, SPSC, SWI, TRIP and VRSK.

Figure 72 – The green group: also characterized by a high percentage of missing data, but a different pattern of intensity in the values. Companies: AYR, CAVN, CTCT, CVLT, GSIT, GUID, IMMR, MNTX, PERI, PRO, RAX, RBCN, TDC, TRAK, TSRA and VOCS, and PRO (medoid).

5.4 Chapter Notes

This chapter presented experimental results of using the proposed interactive visual *TV-MV Analytics* framework to explore time-varying data on ionospheric scintillation and three other application domains. The experimental results show that the multiple coordinated visual representations support diverse exploratory tasks to answer open questions about the data. Experts in ionospheric scintillation provided a positive feedback, as they confirm that *TV-MV Analytics* assists their efforts in further exploring the records on ionospheric scintillation. Also, the representative attributes identified with the assistance of our framework could be monitored using conventional geodesic receivers available, rather than the highly specialized and more costly ISMRs, which implies in significant cost reduction in monitoring the phenomenon.

Although *TV-MV Analytics* framework was designed based on this particular domain, many aspects of its visualization and interaction are generalizable to other fields such as air quality, stock market and crime statistics, when focusing the exploration of time-varying multivariate data. In the results from these application domains, the *TV-MV Analytics* framework showed good potential to answer the speculative questions and scenarios, facilitating the creation of customized analysis pipelines for easy exploration through the time-varying multivariate data sets.

CONCLUSIONS

6.1 Contribution

In this thesis, we introduced the visual analytics framework *TV-MV Analytics* to support exploratory data analysis of complex time-varying multivariate data. The framework incorporates feature extraction, clustering, individual and global visualizations of data variables and data instances, to convey a detailed representation of their temporal behavior at multiple user-defined aggregation levels. It includes overviews of multiple variables regarding their behavioral similarity over user-selected time periods and supports users in identifying representative variables for characterizing a target phenomenon, such as ionospheric scintillation behavior.

We conducted an experimental validation of *TV-MV Analytics*, driven by collected time-varying multivariate data from four different domains: ionospheric scintillation, crime statistics indicators in the state of São Paulo, air quality measurements at an Italian city, and US stock market prices. The case studies are illustrative of how the system can assist experts in gaining further insight into the historical behavior of the multiple domain variables.

The studies with the ionospheric scintillation data were conducted in partnership with domain experts to provide supporting scenarios where *TV-MV Analytics* can be employed to discover relevant information about the scintillation phenomenon. The system was successfully applied to the ionospheric scintillation database to assist the domain experts in answering an illustrative set of relevant questions (Section 5.2), confirming experimentally our initial hypothesis. The investigations indicated that a relatively small subset of the variables tracked [7 – 9] are highly indicative of the behavior of the S4 index and the exploratory analysis relying on visual analytics proved useful to better understand the problem and enable more effective monitoring of the phenomenon.

The design of *TV-MV Analytics* was inspired on the requirements posed by the scintillation data, but many aspects of its visualization and interaction user-driven pipeline can be

generalized to other application domains where time-varying multivariate data arises, such as the ones investigated in Section 5.3. They indicate a good potential of the framework to answer the speculative questions and investigate open scenarios, facilitating the creation of customized exploratory analysis pipelines.

We contribute new strategies for visualizing time-varying multivariate data sets. Visualizations of variables such as (i) a hierarchical similarity map of their feature vectors, (ii) time matrices views for the variables, (iii) correlation matrices from aggregated variables, (iv) time circular diagram, and (v) aggregated and normalized line-graphs. Specialists can use the framework to pose queries relative to a target time period and – assisted by multiple visualizations integrated with analytical algorithms for clustering and classification – can interact with the data and visualizations to find answers to questions and formulate new scientific hypotheses. Our approaches complement existing ones by focusing on the global temporal relationships between the multiple variables tracked, rather than on their individual behavior.

The questions motivating the use cases described were picked as examples illustrative of the potential of the *TV-MV Analytics* framework, but other analysis questions related to these domains could have been addressed, or questions of a similar nature related to other data domains. For instance, Dasgupta et al. (2015) (DASGUPTA; KOSARA; GOSINK, 2015) define analytics questions on time-varying multivariate geological data, to investigate the dynamics of the chemical reactions and understand how they collectively affect the carbon content of the soil. Some of these questions, listed in Table 30, could be investigated with our framework to identify potential hidden patterns in subsets of the data. The *TV-MV Analytics* framework is potentially more useful in situations where: (i) the number of variables is high, (ii) there are defined target variables, and (iii) there is interest in investigating behaviors at distinct temporal scales and time periods.

Table 30 – Questions defined by Dasgupta et al. (2015) (DASGUPTA; KOSARA; GOSINK, 2015)

Which variables are stable or unstable over time?
At which time steps do we see a significant rise or fall in concentration of a variable?
When the concentration of a variable decreases or increases, what is the effect on others?
Which sets of variables show similar levels of concentrations over time?
At which time steps do we see similar concentration levels?
Do the variables behave similarly in particular locations?

The results from early studies of the scintillation database have been presented at a workshop: *Visual Classification of Ionospheric Scintillation Data. Workshop Calibra (MundoGeo # connect - Latin America 2014), São Paulo, Brazil*. In that investigation we combined the LSP projection technique and similarity-based phylogenetic trees to visually investigate global similarities on a subset of ionospheric scintillation data (1,000 data instances). Alternative dissimilarity metrics were considered in the visualizations and classifiers to gain a better understanding of the data set. A description of the design and development of *TV-MV Analytics* and results from

its usage to investigate the scintillation database have been presented as at SIBGRAPI 2017, *The Brazilian Conference on Graphics, Patterns and Images*, and published at the Computers & Graphics journal: Soriano-Vargas, A., Vani, B. C., Shimabukuro, M. H., Monico, J. F., Oliveira, M. C. F., and Hamann, B. *Visual analytics of time-varying multivariate ionospheric scintillation data. Computers & Graphics*, 68, (pp. 96-107). Elsevier 2017 (selected papers from SIBGRAPI).

6.2 Limitations

In developing the case studies of time-varying multivariate data using the proposed framework, the following limitations were identified:

1. Although hierarchical versions of visual representations have been presented, they are still unable to nicely handling cluttering and overusing of space required when exploring a very large number of variables. Exploring many variables with a great level of detail remains a major challenge.
2. *TV-MV Analytics* was designed and implemented bearing in mind the requirements of the ionospheric scintillation experts. Although it can be employed to investigate other data sets of a similar nature, i.e., with multiple variables recorded over extensive time periods, the overall approach adopted in the solution does not necessarily fit to any kind of time-varying multivariate data set, for instance, those with only a few variables.
3. The solution relies on deriving informative feature vectors from the time matrix representations, and several alternatives are provided for doing so. However, the task of selecting the appropriate moments for a given data set is not straightforward, and requires knowledge about the data properties, which may not be available from the start. A similar reasoning applies to the issue of selecting the best dissimilarity measure for the data.
4. A more extensive evaluation of the usability of the time circular diagram (Section 4.6.2) is required, as well as systematic evaluation of the framework with domain experts from different data domains. Such evaluations are important to provide further information on the potential advantages of the proposed solutions and to evolve the framework into a solution truly intuitive to target users (data analytics experts).
5. *TV-MV Analytics* does not handle categorical variables, only time-stamped observations described by numerical variables, though a few categorical attributes may be represented with target classes for visualization purposes.

6.3 Future work

Future work can incorporate new ideas that emerged during our discussions, as well as to advance in handling the aforementioned limitations.

1. Missing data is handled with a linear interpolation function, but it would be interesting to offer other higher-order interpolation schemes and analyze how they impact data analysis (BOYLES, 2011).
2. Further analytical studies to inform the choice of feature moments and dissimilarity functions suitable to time-varying multivariate data exhibiting different properties. Extending the solution to handle data sets with categorical values by defining proper ordering and similarity metrics could also integrate this effort.
3. The use of the space is well explored with our hierarchical strategies. Nevertheless, other techniques may be investigated to improve the visual representations reducing the used space and avoiding the cluttering while preserving information at multiple levels-of-detail.
4. Conduct further validations of the system, including studies with the participation of experts from multiple data domains to gain further insight into the usage of the framework and possible improvements to enhance usability.

BIBLIOGRAPHY

ABO-ZAID, A.; HINTON, O. R.; HORNE, E. About moment normalization and complex moment descriptors. **Pattern Recognition**, Springer, p. 399–409, 1988. Citations on pages [76](#) and [78](#).

ACKAH, J.-B.; OBROU, O.; GROVES, K. Study of the ionospheric scintillation and tec characteristics at solar minimum in a west african equatorial region using global positioning system (gps) data. **30th IEEE General Assembly and Scientific Symposium (URSI)**, p. 1–4, 2011. Citations on pages [16](#) and [60](#).

AKIBAY, H.; MA, K.-L. A tri-space visualization interface for analyzing time-varying multivariate volume data. **Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization**, p. 115–122, 2007. Citation on page [43](#).

ALENCAR, D.; CARVALHO, D.; KOENDERS, E.; MOURÃO, F.; ROCHA, L. Devising a computational model based on data mining techniques to predict concrete compressive strength. **Procedia Computer Science**, Elsevier, v. 108, p. 455–464, 2017. Citation on page [82](#).

ALFONSI, L.; WERNIK, A. W.; MATERASSI, M.; SPOGLI, L.; BOUGARD, B.; MONICO, J. F. G. Low latitude scintillations: a comparison of modeling and observations within the cigala project. **30th IEEE General Assembly and Scientific Symposium (URSI)**, p. 1–4, 2011. Citations on pages [26](#) and [50](#).

ARVATE, P.; SOUZA, A. P. The fire-armed police effect. **New York Times**, 2011. Citations on pages [62](#), [104](#), [110](#), and [113](#).

ASSENT, I. Clustering high dimensional data. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 2, n. 4, p. 340–350, 2012. Citation on page [31](#).

AZIZ, S.; AIN, W.; MAJEED, R.; KHAN, M. A.; QAYUM, I.; AHMED, I.; HOSAIN, K. Growth centile charts (anthropometric measurement) of pakistani pediatric population. **JPMA-Journal of the Pakistan Medical Association**, v. 62, n. 4, p. 367, 2012. Citations on pages [19](#) and [49](#).

BERNARD, J.; DOBERMANN, E.; BÖGL, M.; RÖHLIG, M.; VÖGELE, A.; KOHLHAMMER, J. Visual-interactive segmentation of multivariate time series. **EuroVis Workshop on Visual Analytics (EuroVA)**, p. 31–35, 2016. Citation on page [25](#).

BERNARD, J.; WILHELM, N.; SCHERER, M.; MAY, T.; SCHRECK, T. Timeseriespaths : Projection-based explorative analysis of multivariate time series data. In: **Journal of WSCG**. [S.l.: s.n.], 2012. p. 97–106. ISSN 1213-6972. Citation on page [25](#).

BOTELHO, G. M. **Seleção de características apoiada por mineração visual de dados**. Master's Thesis (Master's Thesis) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011. Citations on pages [36](#) and [38](#).

BOYLES, S. Comparison of interpolation methods for missing traffic volume data. **90th Annual Meeting of Transportation Research Board**, n. 11-3757, 2011. Citation on page [132](#).

CONKER, R. S.; EL-ARINI, M. B.; HEGARTY, C. J.; HSIAO, T. Modeling the effects of ionospheric scintillation on gps/satellite-based augmentation system availability. **Radio Science**, v. 38, n. 1, 2003. Citation on page [55](#).

CRUZ, L. E. F. **Uma abordagem baseada em técnicas de visualização de informações para avaliação de características de imagens e aplicações**. Master's Thesis (Master's Thesis) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012. Citations on pages [36](#) and [38](#).

CUADROS, A. M.; PAULOVICH, F. V.; MINGHIM, R.; TELLES, G. P. Point placement by phylogenetic trees and its application to visual analysis of document collections. **IEEE Symposium on Visual Analytics Science and Technology**, v. 1, p. 99–106, 2007. Citation on page [36](#).

DASGUPTA, A.; KOSARA, R.; GOSINK, L. Vimtex: A visualization interface for multivariate, time-varying, geological data exploration. **Computer Graphics Forum**, v. 34, n. 3, p. 341–350, 2015. Citations on pages [15](#), [20](#), [45](#), [47](#), and [130](#).

DIERENDONCK, A. V.; KLOBUCHAR, J.; HUA, Q. Ionospheric scintillation monitoring using commercial single frequency c/a code receivers. In: **Proceedings of ION GPS**. [S.l.: s.n.], 1993. v. 93, p. 1333–1342. Citation on page [54](#).

DY, J. G.; BRODLEY, C. E. Visualization and interactive feature selection for unsupervised data. **6th Conference of Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)**, ACM, p. 360–364, 2000. Citations on pages [37](#) and [38](#).

EDWARDS, R. D.; BASSETTI, W.; MAGEE, J. **Technical analysis of stock trends**. [S.l.]: CRC press, 2007. Citation on page [64](#).

ENGEL, D.; GREFF, K.; GARTH, C.; BEIN, K.; WEXLER, A.; HAMANN, B.; HAGEN, H. Visual steering and verification of mass spectrometry data factorization in air quality research. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 18, n. 12, p. 2275–2284, 2012. Citation on page [26](#).

FALOUTSOS, C.; LIN, K.-I. **FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets**. [S.l.]: ACM, 1995. Citation on page [89](#).

FRANK, E.; HALL, M.; HOLMES, G.; KIRKBY, R.; PFAHRINGER, B.; WITTEN, I. H.; TRIGG, L. Weka-a machine learning workbench for data mining. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2009. p. 1269–1277. Citation on page [82](#).

GHEYAS, I. A.; SMITH, L. S. Feature subset selection in large dimensionality domains. **Pattern Recognition**, Elsevier, v. 43, n. 1, p. 5–13, 2010. Citation on page [25](#).

HABIBI, M. S.; SHIRKHODAIE, A. A survey of visual analytics for knowledge discovery and content analysis. **Proceedings of SPIE**, p. 8392T, 2012. Citation on page [26](#).

HAN, J. **Data Mining: Concepts and Techniques**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. Citation on page [25](#).

HEGARTY, C.; EL-ARINI, M. B.; KIM, T.; ERICSON, S. Scintillation modeling for gps-wide area augmentation system receivers. **Radio Science**, v. 36, n. 5, p. 1221–1231, 2001. Citations on pages 55 and 56.

HOCHHEISER, H.; SHNEIDERMAN, B. Dynamic query tools for time series data sets: time-box widgets for interactive exploration. **Information Visualization**, v. 3, n. 1, p. 1–18, 2004. Citations on pages 15, 37, 38, 39, and 47.

HU, M.-K. Visual pattern recognition by moment invariants. **Transactions Information Theory, IRE**, v. 8, n. 2, p. 179–187, 1962. Citations on pages 19, 76, and 77.

INSELBERG, A. **Parallel Coordinates: Visual Multidimensional Geometry and Its Applications**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2009. Citation on page 26.

ITUR. Ionospheric propagation data and prediction methods required for the design of satellite services and systems. v. 1, p. 531–12, September 2013. International Telecommunication Union: Recommendation. Available: <<http://www.itu.int/pub/R-REC/en>>. Citation on page 55.

JÄCKLE, D.; FISCHER, F.; SCHRECK, T.; KEIM, D. A. Temporal mds plots for analysis of multivariate data. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 22, n. 1, p. 141–150, 2016. Citations on pages 25 and 31.

JIAO, Y.; MORTON, Y. T.; TAYLOR, S.; PELGRUM, W. Characterization of high-latitude ionospheric scintillation of gps signals. **Radio Science**, Wiley Online Library, v. 48, n. 6, p. 698–708, 2013. Citation on page 73.

JOLLIFFE, I. **Principal component analysis**. [S.l.]: Wiley Online Library, 2002. Citation on page 90.

KAUFMAN, L.; ROUSSEEUW, P. **Clustering by means of medoids**. [S.l.]: North-Holland, 1987. Citation on page 79.

KEIM, D.; ANDRIENKO, G.; FEKETE, J.-D.; GÖRG, C.; KOHLHAMMER, J.; MELANÇON, G. Visual analytics: Definition, process, and challenges. In: **Information visualization**. [S.l.]: Springer, 2008. p. 154–175. Citation on page 26.

KEIM, D. A.; SCHNEIDEWIND, J.; SIPS, M. Circleview: A new approach for visualizing time-related multidimensional data sets. **Working Conference on Advanced Visual Interfaces**, p. 179–182, 2004. Citations on pages 15, 39, 40, and 47.

KENNEDY, C. E.; TURLEY, J. P. Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric icu. **Theoretical Biology and Medical Modelling**, BioMed Central, v. 8, n. 1, p. 40, 2011. Citation on page 25.

KIEFT, P.; AQUINO, M.; DODSON, A. Using ordinary kriging for the creation of scintillation maps. In: . [S.l.: s.n.], 2014. chap. 06. Citation on page 50.

KINTNER, P.; HUMPHREYS, T.; HINKS, J. Gns and ionospheric scintillation. how to survive the next solar maximum. **Inside GNSS**, v. 4, n. 4, p. 22–30, 2009. Citations on pages 15 and 51.

KO, S.; CHO, I.; AFZAL, S.; YAU, C.; CHAE, J.; MALIK, A.; BECK, K.; JANG, Y.; RIBARSKY, W.; EBERT, D. S. A survey on visual analysis approaches for financial data. **Computer Graphics Forum**, Wiley Online Library, v. 35, n. 3, p. 599–617, 2016. Citations on pages 64 and 65.

KOH, H. C.; TAN, G. *et al.* Data mining applications in healthcare. **Journal of healthcare information management**, v. 19, n. 2, p. 65, 2011. Citation on page 82.

KORN, F.; PAGEL, B.-U.; FALOUTSOS, C. On the dimensionality curse and the self-similarity blessing. **IEEE Transactions on Knowledge and Data Engineering**, v. 13, n. 1, p. 96–111, 2001. Citation on page 25.

KUMAR, A. medstarbeamer: A new beamer class. In: **TUG BOAT: communications of the tex users group**. [S.l.: s.n.], 2016. v. 37, n. 1, p. 18–24. Citation on page 80.

KUMAR, N.; LOLLA, V. N.; KEOGH, E. J.; LONARDI, S.; RATANAMAHATANA, C. A. Time-series bitmaps: a practical visualization tool for working with large time series databases. **SDM**, SIAM, p. 531–535, 2005. Citations on pages 15, 41, 42, and 47.

LEE, T.-Y.; SHEN, H.-W. Visualization and exploration of temporal trend relationships in multivariate time-varying data. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1359–1366, 2009. Citation on page 43.

LEI, S.; ZHANG, K. A visual analytics system for financial time-series data. **Proceedings of the 3rd International Symposium on Visual Information Communication**, ACM, p. 20, 2010. Citation on page 65.

LEPIK, Ü.; HEIN, H. Haar wavelets. In: . [S.l.]: Springer, 2014. p. 7–20. Citation on page 88.

LEVKOWITZ, H. **Color theory and modeling for computer graphics, visualization, and multimedia applications**. [S.l.]: Springer Science & Business Media, 1997. Citations on pages 16, 83, 84, and 88.

LIMA, G.; STEPHANY, S.; PAULA, E.; BATISTA, I.; ABDU, M.; REZENDE, L.; AQUINO, M.; DUTRA, A. Correlation analysis between the occurrence of ionospheric scintillation at the magnetic equator and at the southern peak of the equatorial ionization anomaly. **Space Weather**, v. 12, n. 6, p. 406–416, 2014. Citation on page 60.

LIMA, G.; STEPHANY, S.; PAULA, E.; BATISTA, I.; ABDU, M. Prediction of the level of ionospheric scintillation at equatorial latitudes in Brazil using a neural network. **Space Weather**, v. 13, n. 8, p. 446–457, 2015. Citation on page 60.

LIN, J.; KEOGH, E.; LONARDI, S.; CHIU, B. A symbolic representation of time series, with implications for streaming algorithms. **Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery**, p. 2–11, 2003. Citation on page 59.

LIN, J.; KEOGH, E.; LONARDI, S.; LANKFORD, J. P.; NYSTROM, D. M. Visually mining and monitoring massive time series. **Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 460–469, 2004. Citations on pages 15, 40, 41, and 47.

LIU, X.; HU, Y.; NORTH, S.; SHEN, H.-W. Correlated multiples: Spatially coherent small multiples with constrained multi-dimensional scaling. **Computer Graphics Forum**, 2015. Citations on pages 15, 43, and 47.

LIVINGSTON, M. A.; DECKER, J. W.; AI, Z. Evaluating multivariate visualizations on time-varying data. **Conference of Visualization and Data Analysis**, v. 8654, p. 86540N, 2013. Citations on pages 31 and 63.

MACHADO, V.; LEITE, R.; MOURA, F.; CUNHA, S.; SADLO, F.; COMBA, J. L. Visual soccer match analysis using spatiotemporal positions of players. **Computers & Graphics**, v. 68, p. 84 – 95, 2017. Citations on pages [15](#), [45](#), [46](#), [47](#), and [93](#).

MACQUEEN, J. *et al.* Some methods for classification and analysis of multivariate observations. In: **Proceedings of the 5th Berkeley symposium on mathematical statistics and probability**. [S.l.: s.n.], 1967. v. 1, n. 14, p. 281–297. Citation on page [79](#).

MAITRA, S. Moment invariants. **Proceedings of the IEEE**, v. 67, n. 4, p. 697–699, 1979. Citation on page [76](#).

MANIYAR, D. M.; NABNEY, I. T. Data visualization with simultaneous feature selection. **IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB)**, p. 1–8, 2006. Citations on pages [37](#) and [38](#).

MARTIN, S.; QUACH, T.-T. Interactive visualization of multivariate time series data. **International Conference on Augmented Cognition**, p. 322–332, 2016. Citation on page [25](#).

MAY, T.; BANNACH, A.; DAVEY, J.; RUPPERT, T.; KOHLHAMMER, J. Guiding feature subset selection with an interactive visualization. **IEEE Conference on Visual Analytics Science and Technology (VAST)**, p. 111–120, 2011. Citations on pages [35](#) and [38](#).

MINGHIM, R.; PAULOVICH, F. V.; LOPES, A. de A. Content-based text mapping using multi-dimensional projections for exploration of document collections. **Electronic Imaging**, International Society for Optics and Photonics, p. 60600S–60600S, 2006. Citation on page [89](#).

MUNZNER, T. **Visualization analysis and design**. Boca Raton, FL: CRC Press, 2014. (AK Peters Visualization series). Citation on page [26](#).

NG, C. U.; MARTIN, G. R. Automatic selection of attributes by importance in relevance feedback visualisation. **IEEE Proceedings of 8th International Conference on Information Visualisation**, p. 588–595, 2004. Citations on pages [33](#) and [38](#).

NORIEGA, A.; BLANCO, D.; ALVAREZ, B.; GARCIA, A. Dimensional accuracy improvement of fdm square cross-section parts using artificial neural networks and an optimization algorithm. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 69, n. 9-12, p. 2301–2313, 2013. Citation on page [120](#).

OLIVEIRA, M. F. D.; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 9, n. 3, p. 378–394, 2003. Citation on page [26](#).

PAIVA, J. G.; FLORIAN, L.; PEDRINI, H.; TELLES, G.; MINGHIM, R. Improved similarity trees and their application to visual data classification. **IEEE transactions on visualization and computer graphics**, v. 17, n. 12, p. 2459–2468, 2011. Citation on page [79](#).

PAULOVICH, F.; MINGHIM, R. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1229–1236, 2008. Citation on page [90](#).

PAULOVICH, F.; NONATO, L.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, 2008. Citation on page [93](#).

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 14, n. 3, p. 564–575, 2008. Citation on page [37](#).

PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The projection explorer: A flexible tool for projection-based multidimensional visualization. **20th Brazilian Symposium on Computer Graphics and Image Processing**, IEEE, p. 27–36, 2007. Citation on page [89](#).

PELLEG, D.; MOORE, A. W. *et al.* X-means: Extending k-means with efficient estimation of the number of clusters. In: **ICML**. [S.l.: s.n.], 2000. v. 1, p. 727–734. Citation on page [79](#).

PIRINGER, H.; BERGER, W.; HAUSER, H. Quantifying and comparing features in high-dimensional datasets. **12th IEEE International Conference on Information Visualisation**, p. 240–245, 2008. Citations on pages [35](#) and [38](#).

PUPYREV, S.; NACHMANSON, L.; KAUFMANN, M. Improving layered graph layouts with edge bundling. In: SPRINGER. **Graph Drawing**. [S.l.], 2010. v. 6502, p. 329–340. Citation on page [88](#).

RAZENTE, H. L.; CHINO, F. J. T.; BARIONI, M. C. N.; TRAINA, A. J.; TRAINA, C. Visual analysis of feature selection for data mining processes. **SBB**, p. 33–47, 2004. Citation on page [31](#).

REZENDE, L.; PAULA, E. de; STEPHANY, S.; KANTOR, I.; MUELLA, M.; SIQUEIRA, P. de; CORREA, K. Survey and prediction of the ionospheric scintillation using data mining techniques. **Space Weather**, v. 8, n. 6, p. 1–10, 2010. Citation on page [59](#).

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Computational and Applied Mathematics**, v. 20, p. 53 – 65, 1987. Citations on pages [36](#) and [80](#).

RUDOLPH, S.; SAVIKHIN, A.; EBERT, D. S. Finvis: Applied visual analytics for personal financial planning. **Symposium on Visual Analytics Science and Technology (VAST)**, IEEE, p. 195–202, 2009. Citation on page [64](#).

SAMARA, T. **Design elements: A graphic style manual**. [S.l.]: Rockport publishers, 2007. Citation on page [84](#).

SAMET, H. **The design and analysis of spatial data structures**. [S.l.]: Addison-Wesley Reading, MA, 1990. Citations on pages [16](#) and [91](#).

SANTOS, M.; LOPE, J. de *et al.* Orthogonal variant moments features in image analysis. **Information Sciences**, Elsevier, v. 180, n. 6, p. 846–860, 2010. Citations on pages [76](#) and [78](#).

SCHRECK, T.; FELLNER, D.; KEIM, D. Towards automatic feature vector optimization for multimedia applications. **Proceedings of the Symposium on Applied computing**, ACM, p. 1197–1201, 2008. Citations on pages [33](#) and [38](#).

SCHULDT, H. Multi-tier architecture. In: **Encyclopedia of database systems**. [S.l.]: Springer, 2009. p. 1862–1865. Citation on page [70](#).

SEO, J. **Information visualization design for multidimensional data: integrating the rank-by-feature framework with hierarchical clustering**. Phd Thesis (PhD Thesis) — University of Maryland at College Park, 2005. Citations on pages 35 and 38.

SEO, J.; SHNEIDERMAN, B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. **IEEE Symposium on Information Visualization**, p. 65–72, 2004. Citations on pages 35 and 38.

SEO, J.; SHNEIDERMAN, B. Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. **IEEE Transactions on Visualization and Computer Graphics**, v. 12, n. 3, p. 311–322, 2006. Citations on pages 35 and 38.

Septentrio Satellite Navigation. **PolaRxS - Application Manual**. 2.4.1. ed. Belgium, 2013. Citations on pages 19, 20, 52, 53, 54, 56, 57, 143, and 144.

SHARF, S. **The world's largest tech companies 2016: Apple bests Samsung, Microsoft and alphabet**. [S.l.]: Forbes, 2016. Citation on page 119.

SKONE, S.; LACHAPELLE, G.; YAO, D.; YU, W.; WATSON, R. Investigating the impact of ionospheric scintillation using a gps software receiver. **Proceedings of GNSS**, p. 1126–1137, 2005. Citation on page 56.

SORENSEN, E.; BRATH, R. Financial visualization case study: Correlating financial time-series and discrete events to support investment decisions. **17th International Conference Information Visualisation**, IEEE, p. 232–238, 2013. Citation on page 65.

STEED, C. A.; HALSEY, W.; DEHOFF, R.; YODER, S. L.; PAQUIT, V.; POWERS, S. Falcon: Visual analysis of large, irregularly sampled, and multivariate time series data in additive manufacturing. **Computers & Graphics**, Elsevier, v. 63, p. 50–64, 2017. Citations on pages 15, 25, 45, 46, and 47.

STEIGER, M.; BERNARD, J.; MITTELSTÄDT, S.; LÜCKE-TIEKE, H.; KEIM, D.; MAY, T.; KOHLHAMMER, J. Visual analysis of time-series similarities for anomaly detection in sensor networks. **Computer Graphics Forum**, Wiley Online Library, v. 33, n. 3, p. 401–410, 2014. Citations on pages 15, 41, 43, and 47.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. *et al.* A comparison of document clustering techniques. In: BOSTON. **KDD workshop on text mining**. [S.l.], 2000. v. 400, n. 1, p. 525–526. Citation on page 79.

SUN, G.-D.; WU, Y.-C.; LIANG, R.-H.; LIU, S.-X. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. **Journal of Computer Science and Technology**, Springer, v. 28, n. 5, p. 852–867, 2013. Citation on page 26.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multi-dimensional data sets. **Information Visualization**, v. 2, n. 4, p. 218–231, 2003. Citation on page 89.

TIAN, D. ping *et al.* A review on image feature extraction and representation techniques. **International Journal of Multimedia and Ubiquitous Engineering**, v. 8, n. 4, p. 385–396, 2013. Citation on page 76.

TIWARI, R.; SKONE, S.; TIWARI, S.; STRANGWAYS, H. Wbmod assisted pll gps software receiver for mitigating scintillation affect in high latitude region. **30th URSI General Assembly and Scientific Symposium**, p. 1–4, 2011. Citations on pages [55](#) and [56](#).

TODOROVA, L.; VASSILEV, P.; SURCHEV, J. Using phi coefficient to interpret results obtained by intercriteria analysis. In: . [S.l.]: Springer, 2016. p. 231–239. Citation on page [88](#).

TUFTE, E. R. Envisioning information. **Optometry & Vision Science**, v. 68, n. 4, p. 322–324, 1991. Citation on page [43](#).

TURKAY, C.; FILZMOSER, P.; HAUSER, H. Brushing dimensions; a dual visual analysis model for high dimensional data. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 17, n. 12, p. 2591–2599, 2011. Citations on pages [15](#), [34](#), [35](#), [38](#), and [47](#).

TURKAY, C.; LUNDERVOLD, A.; LUNDERVOLD, A. J.; HAUSER, H. Representative factor generation for the interactive visual analysis of high-dimensional data. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 18, n. 12, p. 2621–2630, 2012. Citations on pages [35](#) and [38](#).

TURKAY, C.; PARULEK, J.; REUTER, N.; HAUSER, H. Interactive visual analysis of temporal cluster structures. **Computer Graphics Forum**, Wiley Online Library, v. 30, n. 3, p. 711–720, 2011. Citations on pages [15](#), [40](#), [42](#), and [47](#).

VANI, B.; MONICO, J. G.; SHIMABUKURO, M.; PEREIRA, V.; AQUINO, M. Exploring the cigala/calibra network data base for ionosphere monitoring over Brazil. **AGU Fall Meeting Abstracts**, v. 1, 2013. Citations on pages [15](#), [27](#), [50](#), [51](#), [54](#), and [57](#).

VANI, B. C. **Análise da cintilação ionosférica no Brasil empregando GNSS e técnicas de mineração e visualização de dados**. Master's Thesis (Master's Thesis) — Universidade Estadual Paulista (UNESP), 2014. Citations on pages [15](#), [16](#), [26](#), [27](#), [52](#), [57](#), [58](#), [59](#), and [60](#).

VITO, S. D.; MASSERA, E.; PIGA, M.; MARTINOTTO, L.; FRANCA, G. D. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. **Sensors and Actuators B: Chemical**, Elsevier, v. 129, n. 2, p. 750–757, 2008. Citations on pages [19](#), [20](#), [63](#), [116](#), and [117](#).

VITO, S. D.; PIGA, M.; MARTINOTTO, L.; FRANCA, G. D. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. **Sensors and Actuators B: Chemical**, Elsevier, v. 143, n. 1, p. 182–191, 2009. Citations on pages [63](#) and [64](#).

WIJK, J. J. V.; SELOW, E. R. V. Cluster and calendar based visualization of time series data. **IEEE Symposium on Information Visualization**, p. 4–9, 1999. Citations on pages [15](#), [39](#), [40](#), and [47](#).

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016. Citation on page [85](#).

YANG, H. H.; MOODY, J. Data visualization and feature selection: New algorithms for nongaussian data. **Advances in Neural Information Processing Systems**, p. 687–693, 2000. Citations on pages [31](#), [36](#), and [38](#).

- YANG, J.; HUBBALL, D.; WARD, M. O.; RUNDENSTEINER, E. A.; RIBARSKY, W. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 13, n. 3, p. 494–507, 2007. Citations on pages [32](#), [33](#), and [38](#).
- YANG, J.; PENG, W.; WARD, M. O.; RUNDENSTEINER, E. A. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. **IEEE Symposium on Information Visualization (INFOVIS)**, p. 105–112, 2003. Citations on pages [32](#) and [38](#).
- YANG, J.; WARD, M. O.; RUNDENSTEINER, E. A.; HUANG, S. Visual hierarchical dimension reduction for exploration of high dimensional datasets. **Proceedings of the Symposium on Data Visualisation (VISSYM)**, Eurographics Association, p. 19–28, 2003. Citations on pages [15](#), [32](#), and [38](#).
- YAZDANBAKHS, O.; DICK, S. Forecasting of multivariate time series via complex fuzzy logic. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 47, n. 8, p. 2160–2171, 2017. Citation on page [25](#).
- YEH, K.; LIU, C. Radio wave scintillations in the ionosphere. **Proceedings of IEEE**, v. 70, n. 4, p. 324–360, 1982. Citation on page [50](#).
- YI, B.-K.; FALOUTSOS, C. Fast time sequence indexing for arbitrary lp norms. **Proceedings of the 26th International Conference on Very Large Data Bases**, p. 385–394, 2000. Citation on page [58](#).
- YUAN, X.; REN, D.; WANG, Z.; GUO, C. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. **IEEE Transactions on Visualization and Computer Graphics (TVCG)**, v. 19, n. 12, p. 2625–2633, 2013. Citations on pages [15](#), [33](#), [34](#), [38](#), and [47](#).
- ZHANG, D.; ZHU, L.; WANG, C.; ZHANG, L. Timespiral, an enhanced interactive visual system for time series data. **2nd International Conference on Information Management (ICIM)**, p. 127–133, 2016. Citations on pages [44](#) and [47](#).
- ZHANG, J.; AHLBRAND, B.; MALIK, A.; CHAE, J.; MIN, Z.; KO, S.; EBERT, D. S. A visual analytics framework for microblog data analysis at multiple scales of aggregation. **Computer Graphics Forum**, v. 35, n. 3, p. 441–450, 2016. Citations on pages [26](#) and [36](#).
- ZHANG, J.; GRUENWALD, L. Opening the black box of feature extraction: incorporating visualization into high-dimensional data mining processes. **6th IEEE International Conference on Data Mining**, p. 1188–1192, 2006. Citation on page [38](#).
- ZIEGLER, H.; JENNY, M.; GRUSE, T.; KEIM, D. A. Visual market sector analysis for financial time series data. **Symposium on Visual Analytics Science and Technology (VAST)**, IEEE, p. 83–90, 2010. Citations on pages [16](#) and [65](#).

IONOSPHERIC SCINTILLATION VARIABLES

Tables 31 and 32 present the variables descriptive of the behavior of ionospheric scintillation. The gray rows indicate identifiers, from which the svid is used as source and the other ones are not considered in our data analysis because they are not numeric values. The remaining variables can be categorized in several types, as described next.

Table 31 – Variables that describe the behavior of ionospheric scintillation (Septentrio Satellite Navigation, 2013) (Part 1). The gray rows refer to identifiers.

#	Variable	Description	Category
	<i>svid</i>	satellite id	identifier
	<i>sbfd_block</i>	Value of RxState of the Receiver Status	identifier
	<i>sbfd2ismr</i>	sbfd2ismr version number	identifier
1	<i>avg_cn0_l1</i>	Average L1 C/N0 over last minute	index
2	<i>avg_cn0_l2</i>	Average L2 C/N0 over the last minute	index
3	<i>avg_cn0_l5_e5b</i>	Average L3 C/N0 over the last minute	index
4	<i>avgccd_l1</i>	average of code/carrier divergence	index
5	<i>avgccd_l2c_e5a</i>	average of code/carrier divergence on L2	index
6	<i>avgccd_l5_e5b</i>	average of code/carrier divergence on L3	index
7	<i>sigmaccd_l1</i>	standard deviation of code/carrier divergence	index
8	<i>sigmaccd_l2c_e5a</i>	standard deviation of code/carrier divergence on L2	index
9	<i>sigmaccd_l5_e5b</i>	standard deviation of code/carrier divergence on L3	index
10	<i>f2nd_tec_cn0</i>	Averaged C/N0 of 2nd frequency	index
11	<i>tec</i>	TEC at TOW	index
12	<i>tec_l5</i>	TEC at TOW - 15 seconds	index
13	<i>tec_30</i>	TEC at TOW - 30 seconds	index
14	<i>tec_45</i>	TEC at TOW - 45 seconds	index
15	<i>dtec_l5tow</i>	dTEC from TOW - 15s to TOW	index
16	<i>dtec_3015</i>	dTEC from TOW - 30s to TOW - 15s	index
17	<i>dtec_4530</i>	dTEC from TOW - 45s to TOW - 30s	index
18	<i>dtec_6045</i>	dTEC from TOW - 60s to TOW - 45s	index
19	<i>phi01l1</i>	Phi01 on L1 - 1-second phase sigma (radians)	index
20	<i>phi03l1</i>	Phi03 on L1 - 3-second phase sigma	index
21	<i>phi10l1</i>	Phi10 on L1 - 10-second phase sigma	index
22	<i>phi30l1</i>	Phi30 on L1 - 30-second phase sigma	index
23	<i>phi60l1</i>	Phi60 on L1 - 60-second phase sigma	index
24	<i>phi01_l2c_e5a</i>	Phi01 on L2-1-second phase sigma	index
25	<i>phi03_l2c_e5a</i>	Phi03 on L2-3-second phase sigma	index
26	<i>phi10_l2c_e5a</i>	Phi10 on L2 - 10-second phase sigma	index
27	<i>phi30_l2c_e5a</i>	Phi30 on L2 - 30-second phase sigma	index
28	<i>phi60_l2c_e5a</i>	Phi60 on L2 - 60-second phase sigma	index
29	<i>phi01_l5_e5b</i>	Phi01 on L3 - 1-second phase sigma	index
30	<i>phi03_l5_e5b</i>	Phi03 on L3 - 3-second phase sigma	index
31	<i>phi10_l5_e5b</i>	Phi10 on L3 - 10-second phase sigma	index

Table 32 – Variables that describe the behavior of ionospheric scintillation ([Septentrio Satellite Navigation, 2013](#)) (Part 2). The gray rows refer to identifiers.

#	Variable	Description	Category
32	<i>phi30_l5_e5b</i>	Phi30 on L3 - 30-second phase sigma	index
33	<i>phi60_l5_e5b</i>	Phi60 on L3 - 60-second phase sigma	index
34	<i>p_l1</i>	P(spectral slope of detrended phase) on L1	index
35	<i>p_l2c_e5a</i>	P(phase spectral slope) on L2	index
36	<i>p_l5_e5b</i>	P (spectral slope of detrended phase)on L3	index
37	<i>t_l1</i>	T (phase power spectral density) on L1	index
38	<i>t_l2c_e5a</i>	T (phase power spectral density) on L2	index
39	<i>t_e5b</i>	T (phase power spectral density) on L3	index
40	<i>vphi60_l1</i>	Verticalization of Phi60 on L1	index
41	<i>vphi60_l2c_e5a</i>	Verticalization of Phi60 on L2	index
42	<i>vphi60_l5_e5b</i>	Verticalization of Phi60 on L3	index
43	<i>vtec</i>	Verticalization of tec	index
44	<i>s4</i>	Total S4 on L1	s4_index
45	<i>s4_l2</i>	Total S4 on L2	s4_index
46	<i>s4_l5_e5b</i>	Total S4 on L3 (dimensionless)	s4_index
47	<i>corr_s4</i>	Correction to total S4 on L1	s4_index
48	<i>corr_s4_l2</i>	Correction to total S4 on L2	s4_index
49	<i>corr_s4_l5_e5b</i>	Correction to total S4 on L3	s4_index
50	<i>s4_corrected</i>	Total S4 on L1 corrected	s4_index
51	<i>s4_l2_corrected</i>	Total S4 on L2 corrected	s4_index
52	<i>s4_l5_e5b_corrected</i>	Total S4 on L3 corrected	s4_index
53	<i>vs4_l1</i>	Verticalization of S4 on L1	s4_index
54	<i>vs4_l2</i>	Verticalization of S4 on L2	s4_index
55	<i>vs4_l5</i>	Verticalization of S4 on L3	s4_index
56	<i>si_l1</i>	SI Index on L1	si_index
57	<i>si_l2c_e5a</i>	SI Index on L2	si_index
58	<i>si_l5_e5b</i>	SI Index on L3	si_index
59	<i>numerator_si_l1</i>	SI Index on L1 - numerator only	si_index
60	<i>numerator_si_l2c_e5a</i>	SI Index on L2, numerator only	si_index
61	<i>numerator_si_l5_e5b</i>	SI Index on L3, numerator only	si_index
62	<i>wn</i>	GPS week number	time
63	<i>tow</i>	GPS time of week	time
64	<i>dow</i>	Day of week	time
65	<i>f2nd_tec_locktime</i>	Lock time on 2nd frequency	time
66	<i>l1_locktime</i>	L1 lock time (seconds)	time
67	<i>l2_e5a_locktime</i>	L2 lock time	time
68	<i>l5_e5b_locktime</i>	L3 lock time	time
69	<i>azim</i>	Azimuth	direction
70	<i>elev</i>	Elevation	direction

XML CONFIGURATION FILE

In view of the sizes of data sets handled, database management systems must be employed for real-time data manipulation. For that reason, the architecture of the framework includes an XML layer which is basically a configuration file to enable it to connect to databases from multiple domains. This strategy provides an effortless approach to handle data from diverse domains: in this work, databases and XML files were instantiated with four distinct data sets.

Figure 73 illustrates a basic example of an XML configuration file structure. The designed structure of the XML configuration file includes diverse tags related to:

1. Connection: host, driver, prefix, port, name of the database, user and password.
2. Tables names and elements: Table *variable* (name, id field, name field and category field). Table *source1*, which is the source of instances (name, id field, name field, description field and category field). An optional Table *source2* that can be empty, necessary in cases such as scintillation where satellites (*source1*) belong to stations (*source2*). And the table *instance* (name, id field, source1 field, source2 field (optional), time field, the threshold attribute and options for that threshold (optional, required in the scintillation database, in which attribute *elevation* is employed as a data threshold).
3. Interface settings: the row unit (for time matrices), the time unit for aggregation (can be more of one: minute, day, month and/or year), the aggregation functions available (also can be more of one: average, maximum, minimum, standard deviation and/or median), the PL/SQL function names for each aggregation option, and the categorizations (also can be more of one) detailing the variable used as target and the value ranges.

Figure 73 – Basic example of an XML configuration file structure.