UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Mining of rainfall patterns from social media for supporting flood risk management

Sidgley Camargo de Andrade

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: _____

Sidgley Camargo de Andrade

Mining of rainfall patterns from social media for supporting flood risk management

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Alexandre Cláudio Botazzo Delbem

Co-advisor: Prof. Dr. João Porto de Albuquerque Pereira

USP – São Carlos
June 2020

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

A553m	<p>Andrade, Sidgley Camargo de Mining of rainfall patterns from social media for supporting flood risk management / Sidgley Camargo de Andrade; orientador Alexandre Cláudio Botazzo Delbem; coorientador João Porto de Albuquerque Pereira. -- São Carlos, 2020. 113 p.</p> <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, 2020.</p> <p>1. Spatio-temporal analysis. 2. Spatial data mining. 3. Social media. 4. Rain patterns. 5. Flood risk management. I. Delbem, Alexandre Cláudio Botazzo, orient. II. Pereira, João Porto de Albuquerque, coorient. III. Título.</p>
-------	--

Sidgley Camargo de Andrade

**Mineração de padrões de chuvas nas redes sociais para
apoiar a gestão de risco de inundação**

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

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Alexandre Cláudio Botazzo Delbem

Coorientador: Prof. Dr. João Porto de Albuquerque Pereira

**USP – São Carlos
Junho de 2020**

ACKNOWLEDGEMENTS

First, I am thankful to my family. My wife, Elaine Teixeira Marra, and my parents, Vânia Camargo de Andrade and Reinaldo Bernardin de Andrade, who walked with me through this PhD journey, believed in my potential and gave me the incentive to keep going. I will always be thankful for everything.

I am also truly grateful to my advisor and co-advisor, Dr. Alexandre Cláudio Botazzo Delbem and Dr. João Porto de Albuquerque Pereira, respectively, for all your teachings throughout this PhD. Without their guidance, I would not have become the researcher I am today. I will always be indebted to them for their encouragement and support since the moment I expressed an interest in studying a PhD at the University of São Paulo.

I would also like to thank all the professors, researchers, experts, and staff I had the pleasure of meeting throughout this PhD journey, most notably the committee members for reading this thesis and making many suggestions which have been of the greatest service to my work. Many thanks to Dr. Leonardo Bacelar Lima Santos, Dr. Cláudio Elízio Calazans Campelo, and Dr. Carlos Dias Maciel.

I am also thankful to my friends and colleagues at the Institute of Mathematics and Computer Sciences (ICMC) and the São Carlos School of Engineering (EESC), both of the University of São Paulo (USP), for providing a nice and friendly workplace environment. Each one of you contributed to where I am today. Special thanks to my friends Camilo Restrepo-Estrada and Livia Castro Degrossi for always being willing to discuss ideas, provide advice, and collaborate with my research.

I would also like to thank the Federal University of Technology – Paraná for supporting me with working time discharging, and the Center for Mathematical Sciences Applied to Industry (CeMEAI) (grant number #2013/07375-0), funded by São Paulo Research Foundation (FAPESP), for providing the necessary computing resources – the Euler cluster is wonderful.

And last but not least, I also would like to acknowledge the financial support provided by the São Paulo Research Foundation (FAPESP) (grant numbers #2017/15413-0 and #2019/01717-2), Coordination for the Improvement of Higher Education Personnel (CAPES) (grant numbers #88882.328783/2014-01 (PROEX) and #88887.091744/2014-01 (Pró-Alertas)), Araucária Foundation of Support for the Scientific and Technological Development of the State of Paraná (FAPPR) (CP 18/2015), and Superintendence of Science, Technology and Higher Education of the State of Paraná (SETI). Such financial supports were essential for the processes of developing and publishing this research, as well as in getting this PhD's degree.

*“Everything is in a state of flux.
You can’t step in the same river twice.”
(Heraclitus)*

RESUMO

DE ANDRADE, S. C. **Mineração de padrões de chuvas nas redes sociais para apoiar a gestão de risco de inundação**. 2020. 113 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, 2020.

Contexto. O uso generalizado de plataformas de rede social e telefones celulares nos últimos anos tem aumentado a capacidade das pessoas de compartilhar informações a qualquer hora, em qualquer lugar, e sobre qualquer tópico. Os últimos anos testemunharam um interesse crescente em dados de rede social como uma fonte suplementar para a gestão de risco de desastres. A maioria dos estudos teve como objetivo extrair padrões temáticos espaço-temporais das redes sociais para apoiar as tarefas de gestão de risco de desastres. Avanços foram feitos no entendimento de padrões temáticos espaço-temporais de fenômenos naturais, tais como padrões de inundações e terremotos. **Lacuna.** No entanto, pouca atenção foi dada aos padrões de chuva, que são entradas fundamentais em muitos modelos chuva-vazão para a modelagem e previsão de inundação, bem como para sistemas de alerta precoce de condições meteorológicas extremas. Questões como a seleção de uma unidade de agregação de área representativa, validação/calibração temporal com dados convencionais, e melhoria do processo de recuperação de informação não foram investigadas exaustivamente e ainda podem ser levantadas como desafios para o estabelecimento de sinais sociais mais sofisticados que são capazes de refletir fenômenos naturais. **Contribuição.** Esta tese de doutorado contribui para a extração de padrões de chuva dos dados do Twitter para apoiar o monitoramento e a previsão de riscos de inundação. Também avança no estabelecimento de (i) um método sistemático para a seleção de uma unidade de área ideal, (ii) uma abordagem para a avaliação da validade temporal da atividade de rede social relacionada a um determinado fenômeno de interesse, (iii) um modelo conceitual para caracterizar as unidades espaciais em que o sinal social espelha com precisão um determinado fenômeno de interesse, e (iv) uma análise de sensibilidade dos padrões espaço-temporais de palavras-chaves relacionadas ao fenômeno de interesse. Uma série de estudos de caso foram conduzidos na cidade de São Paulo, Brasil, a fim de avaliar as contribuições. **Resultados.** Os resultados mostraram a viabilidade de extrair padrões de chuva dos dados do Twitter e seu uso na tolerância a falhas de soluções tradicionais de gestão de risco de inundação, especialmente em áreas onde há ausência de dados convencionais. **Conclusões.** Os dados de redes sociais podem ser usados como uma fonte de dados suplementar para monitoramento de chuvas. Além disso, discussões fornecem princípios orientadores úteis a serem seguidos por analistas espaciais ao usar dados de redes sociais como uma fonte de dados proxy de fenômenos naturais.

Palavras-chave: Análise espaço-temporal, Mineração de dados espaciais, Redes sociais, Padrões de chuva, Gestão de risco de inundação.

ABSTRACT

DE ANDRADE, S. C. **Mining of rainfall patterns from social media for supporting flood risk management.** 2020. 113 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, 2020.

Context. The widespread use of social media platforms and mobile phones in recent years has increased the capability of people to share information anytime, anywhere, and about anything. The past few years have witnessed a growing interest in social media data as a supplementary data source for disaster risk management. Most studies have aimed at extracting spatio-temporal thematic patterns from social media to support the wide range of tasks that comprise disaster risk management. Substantial advances have been made towards the understanding patterns of several natural phenomena, such as floods and earthquakes. **Gap.** However, scant attention has been given to rain patterns, which are fundamental inputs in many rainfall-runoff models for flood modeling and forecasting, as well as early warning systems of extreme weather. Furthermore, issues such as selection of a representative areal unit of aggregation, temporal validation/calibration with conventional data, and improvement in information retrieval processes have not been thoroughly investigated, and can still be raised as challenges for the establishment of more sophisticated social signals that reflect natural phenomena. **Contribution.** This doctoral thesis contributes to the extraction of rain patterns from Twitter data for supporting monitoring and forecasting in flood risk management. It advances in establishing (i) a systematic method for the selection of an optimal areal unit, (ii) an approach for the evaluation of the temporal validity of social media activity related to a given phenomenon of interest, (iii) a conceptual specification model for characterization of the spatial units where social signals accurately mirror a given phenomenon of interest, and (iv) a sensitivity analysis of the spatio-temporal patterns of keywords related to a given phenomenon of interest. A series of empirical case studies conducted in Sao Paulo city, Brazil, evaluated such contributions. **Results.** The results showed the viability of extraction of rain patterns from Twitter data and their potential use to improve the fault tolerance of traditional solutions of flood risk management, especially in areas of lack of conventional data. **Conclusions.** Social media data can be used as a supplementary data source for rainfall monitoring. Moreover, discussions have provided useful guiding principles to be followed by spatial analysts using social media data as a proxy data source of natural phenomena.

Keywords: Spatio-temporal analysis, Spatial data mining, Social media, Rain patterns, Flood risk management.

LIST OF FIGURES

Figure 1 – Basic indicators of the International Report on Losses in Brazil from 1990 to 2014.	26
Figure 2 – Scale and zoning effects for spatial data aggregation. (a) corresponds to a sample of spatial data at an individual level bounded to a study area, (b) is the spatial count data using a reference lattice design, (c) and (d) illustrate the scale and zoning effects, respectively.	35
Figure 3 – Trade-off between Global Moran's I and the overall degree of structural (in)stability (standard deviation) of Local Moran's I . It should be noted the difference when compared the trends of the standard deviation of Local Moran's I and Global Moran's I across the lattices. Standard deviation of the Local Moran's I was normalized by scaling between the minimum and maximum values of the Global Moran's I . Both statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.	37
Figure 4 – Methodological multicriteria optimization framework for the selection of an optimal areal unit in a spatial data analysis.	38
Figure 5 – Bootstrap resampling strategy with set of spatial data grouped into regular time units called events. • corresponds to spatial data at an individual level spread over a study area (e.g., geotagged social media messages across the city of Sao Paulo).	41
Figure 6 – Cross-section data of daily rainfall and frequency of rain-related geotagged tweets from 7 November 2016 to 26 April 2017, Sao Paulo, Brazil.	44
Figure 7 – Trade-off between the global indicator of spatial association (Global Moran's I) and the overall degree of structural (in)stability (coefficient of variation of Local Moran's I normalized by scaling between the minimum and maximum values of the Global Moran's I coefficients. Both global and local spatial statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.	44
Figure 8 – Pareto frontier and trade-off between Global Moran's I and the coefficient of variation of Local Moran's I (overall degree of structural (in)stability). Both statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.	45

Figure 9 – Robustness of the Pareto-optimal areal units using the bootstrap method with 1,000 replications – for details of the bootstrap resampling strategy, see Section 2.2.	46
Figure 10 – Comparison of spatial patterns of Pareto-optimal areal units (middle) and four arbitrary areal units (extremes). The patterns correspond to the ‘odds ratio measure’ of the frequency of geotagged tweets (POORTHUIS <i>et al.</i> , 2014). ℓ corresponds to the side length of hexagonal lattices.	47
Figure 11 – Increasing rainfall and frequency of rain-related geotagged tweets from 1 14:00 BRST to 3 00:00 BRST January 2016, Sao Paulo, Brazil (10-min temporal scale).	55
Figure 12 – Increasing rainfall and frequency of rain-related geotagged tweets from 25 12:00 BRST to 28 08:00 BRST January 2016, Sao Paulo, Brazil (10-min temporal scale).	55
Figure 13 – Map of the distribution of the rain-related geotagged tweets vis-à-vis the rain gauges and topographic data, in Sao Paulo, Brazil.	57
Figure 14 – Methodological approach for designing the temporal approach.	59
Figure 15 – Spatial interpolation of the active rain gauges in the city of Sao Paulo, Brazil.	61
Figure 16 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 14:00 BRST to 3 00:00 BRST January 2016.	62
Figure 17 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 25 12:00 BRST to 28 08:00 BRST January 2016.	63
Figure 18 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 to 30 January 2016.	65
Figure 19 – Cross-correlation and visualization between the times series of rainfall data and rain-related geotagged tweets with a temporal scale of 20-min for the period from 1 to 30 January 2016.	66
Figure 20 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 30-min for the period from 1 to 30 January 2016.	67
Figure 21 – Comparison of cross-correlation between the time series of rainfall data and rain-related geotagged tweets with temporal scales of 10, 20 and 30-min for the period from 1 to 30 January 2016.	68
Figure 22 – (a) Aggregation of geotagged tweets into districts of the city of Sao Paulo and (b) the power law distribution of geotagged tweets. Both for the period from 7 to 30 November 2017.	73

Figure 23 – Example of reshaping between two arbitrary and incompatible zoning systems. (a) two incompatible zoning systems (1–4 and A,B). (b) the original inflow/outflow matrix (or Origin-Destination matrix). (c) the weight matrix based on the overlap between the areas of the zoning systems. (d) the overlapping equation. (e) the calculation of the aggregation vector containing the number of inflows for each area of the reshaped zoning system.	77
Figure 24 – Quantile maps of the (a) population, (b) number of tweets by population, (c) income, and (d) well-educated.	78
Figure 25 – Quantile maps of the (a) intensity of the social mirroring of rain patterns, (b) population, (c) income, and (d) inflow.	79
Figure 26 – Residual map of the conceptual specification model (Equation 4.1).	81
Figure 27 – Scatterplot between the intensity of social mirroring of rainfall patterns and the independent variables of (a) population, (b) income, and (c) inflow. . . .	82
Figure 28 – Choropleth map of the number of filtered tweets per district of the city of Sao Paulo.	89
Figure 29 – Hovmöller-based diagram depicting the signal (1.0) and noise (-1.0) of the keywords over five temporal scales of aggregation and across four highlighted districts (see Figure 28). The x and y axes show the temporal scales of aggregation and the keywords, respectively. The blue color represents the signal intensity, whereas the red color represents the noise intensity. White color represents no data. The signal and noise were measured as the fraction between on-topic and off-topic geotagged tweets and all the geotagged tweets posted within the district (relative frequency) and, later, rescaled to [-1, 1]. . .	91
Figure 30 – Illustrations of the degree of spatial autocorrelation.	111
Figure 31 – Illustration of the two major axes of spatial heterogeneity: compositional and configurational heterogeneity. Each large square is an area and different colors represent different ranges of values (or quantiles). Compositional heterogeneity increases as values/areas belong to different intervals of values (or quantiles). Configurational heterogeneity increases with the increasing complexity of the spatial arrangement.	113

LIST OF CHARTS

Chart 1 – Keywords in Brazilian-Portuguese with their English meaning in parentheses. The keywords were chosen based on previous work (ANDRADE <i>et al.</i> , 2017) and a preliminary analysis of the Twitter messages. Similar terms as “chuva” (rain) and “chuvaaa” (rainn) were aggregated. Keywords with grammar mistakes were take into account as long as the frequency was equal or greater than 10.	89
--	----

LIST OF TABLES

Table 1 – Criteria values and the corresponding Pareto frontier for each areal unit. In the case of Global Moran’s I criterion, the p -value corresponds to the null hypothesis that the social media pattern is randomly distributed among the spatial units.	46
Table 2 – Pareto-optimal areal units and corresponding frequencies after the bootstrap method with 1,000 replications.	47
Table 3 – Examples of geotagged tweets related to the rain phenomenon and classified per time-span (i.e., before, after and during the rain phenomenon).	54
Table 4 – Examples of rain-related geotagged tweets that have been classified manually.	58
Table 5 – Cross-correlation values between rainfall data time series and rain-related tweets with a temporal scale of 10-min for the period from 1 14:00 BRST to 3 00:00 BRST January 2016.	62
Table 6 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 25 12:00 BRST to 28 08:00 BRST January 2016.	64
Table 7 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 to 30 January 2016.	64
Table 8 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 20-min for the period from 1 to 30 January 2016.	66
Table 9 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 30-min for the period from 1 to 30 January 2016.	67
Table 10 – OLS regression statistic results.	81

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
CEMADEN	Brazilian National Center of Monitoring and Early Warning of Natural Disaster
DAEE	Department of Water and Power
FCTH	Hydraulics Technology Foundation Center
GIS	Geographic Information Systems
GIScience	Geographic Information Science
IBGE	Brazilian Institute of Geography and Statistics
IDW	Inverse Distance Weighting
LM test	Lagrange Multiplier test
MAUP	Modifiable Areal Unit Problem
MCDA	Multi-Criteria Decision Analysis
PDM	Probability Distributed Model
SAR	Spatial Autoregressive Regression
SER	Spatial Error Regression
SPMR	Metro Subway Origin & Destination Survey for the Sao Paulo Metropolitan Region
SRI	Surface Rain Intensity
USGS	U.S. Geological Survey

CONTENTS

1	INTRODUCTION	25
1.1	Contextualization	26
1.2	Challenges and goals	28
1.3	Contributions	30
1.4	Thesis outline	32
2	A MULTICRITERIA OPTIMIZATION FRAMEWORK	33
2.1	MAUP in social media analysis	34
2.2	Multicriteria optimization framework	38
2.3	Application of the multicriteria optimization framework	41
2.4	Results	43
2.5	Discussion	48
2.6	Final remarks	50
3	A TEMPORAL VALIDITY APPROACH	53
3.1	The temporal validity problem	54
3.2	Case study	56
3.3	Methodology	58
3.4	Results	61
3.5	Discussion	68
3.6	Final remarks	69
4	A CONCEPTUAL SPECIFICATION MODEL	71
4.1	The spatial heterogeneity in social media data	72
4.2	Materials and methods	74
4.3	The conceptual specification model	76
4.4	Results	79
4.5	Discussion	83
4.6	Final remarks	84
5	A KEYWORD SENSITIVITY ANALYSIS	87
5.1	Case study and methodology	88
5.2	Results	90
5.3	Discussion	93

5.4 Final remarks 94

6 CONCLUSIONS 95

6.1 Contributions 96

6.2 Limitations 97

6.3 Recommendations for future work 98

6.4 Data and codes availability statement 99

BIBLIOGRAPHY 101

APPENDIX A THEORETICAL BACKGROUND 109

A.1 Geographic information science 109

A.2 Spatial data analysis 109

INTRODUCTION

The purpose of this doctoral research is the mining of spatio-temporal thematic patterns from social media activity for the establishment of social signals that reflect natural phenomena, in particular, the social mirroring of heavy rainfall patterns. In Brazil, especially in large urban centers, heavy rain has very often caused flash floods, inundations, and flooding, mainly due to rivers overflowing and poor drainage system of urban pavements. Such a type of disaster has motivated the development of this research, since it has caused human and material losses, economic disruption, and social problems in many cities ([UNISDR, 2015](#)). [Figure 1](#) shows the overall impact of flooding in Brazil from the basic indicators of frequency, mortality, and economic issues available in the International Disaster Database ([CRED EM-DAT, 2015](#)). As can be seen, flooding is the most frequent disaster, responsible for 82% of deaths and 34% of economic losses. Until 2014, in Brazil, more than 7,404 deaths had occurred and 15,077,504 people had been affected by floods and landslides caused largely by heavy rainfall ([ARAGÓN-DURAND, 2014](#)).

The contributions reported in Chapters [2](#), [3](#), [4](#), and [5](#) are original to the field of computer science and bear scientific relevance to the so-called Geographic Information Science (GIScience) – which examines the nature of geographic information and the development and use of theories, methods, technologies, and data for understanding geographic processes, relationships, and patterns on different geographical scales ([MARK, 2003](#); [GOODCHILD, 1992](#)) – and social and economic relevance in supporting the construction of rainfall-runoff models for flood risk management from using social media data. Such models aim to reduce the likelihood and/or impact of floods, nonetheless, their application for flood risk management using social media data is outside the scope of this research and, can be further introduced by [Restrepo-Estrada *et al.* \(2018\)](#) and [Restrepo-Estrada \(2018\)](#).

Figure 1 – Basic indicators of the International Report on Losses in Brazil from 1990 to 2014.

Note – Graphics available at <https://www.preventionweb.net/countries/bra/data/> and accessed on December 29th, 2019.

1.1 Contextualization

Over the past few years, researchers have used public social media data as a data source to study several types of human activities and physical phenomena. Given the widespread use of social media in cities, the analysis of social media activity is promising in the emerging field of urban analytics (SINGLETON; SPIELMAN; FOLCH, 2018). Geospatial data extracted from social media can give an insight about the dynamic patterns of urban environments and urban life in higher spatial and temporal resolutions than has so far been made possible by conventional data sources (e.g., census data and field surveys) (BATTY, 2013). In line with this approach, researchers have used georeferenced social media data to study several key areas, such as detection, monitoring and recognition of natural disasters (e.g., floods, earthquakes, and hurricanes) and humanitarian crises (e.g., outbreaks of epidemic diseases), as well as to tackle urban planning problems (e.g., traffic jams, human mobility, and placement of police patrol routes) – for an overview of this literature, see Martí, Serrano-Estrada and Nolasco-Cirugeda

(2019), Martínez-Rojas, Pardo-Ferreira and Rubio-Romero (2018), Nummi (2017), Manca *et al.* (2017), and Steiger, Albuquerque and Zipf (2015).

A common strategy employed in the previous literature is to assess the intensity of social media activity around a topic and use it as a proxy signal that reveals the spatio-temporal distribution of a phenomenon of interest, i.e., those studies assume the existence of a correlation between the aggregated thematic social media activity in an areal and temporal unit and a given spatio-temporal process. This approach has proved fruitful for the study of natural phenomena, such as flooding (ROSSER; LEIBOVICI; JACKSON, 2017; LI *et al.*, 2018; ARTHUR *et al.*, 2018; SMITH *et al.*, 2017; ALBUQUERQUE *et al.*, 2015), earthquakes (EARLE; BOWDEN; GUY, 2012; SAKAKI; OKAZAKI; MATSUO, 2010), and hurricanes (KRYVASHEYEU *et al.*, 2016), as well as social processes, such as geodemographic patterns (PATEL *et al.*, 2017; LONGLEY; ADNAN, 2016; STEIGER *et al.*, 2015).

In the establishment of a spatio-temporal thematic correlation between social media activity and a given natural phenomenon of interest, phenomenon-related information shared by people through social media platforms anytime and at different points in cities can be collectively valuable as a supplementary data source to conventional data sources (e.g., physical sensors). Although considered an “opportunistic sensing” (i.e., when social media users are not aware their messages are part of a spatio-temporal thematic mining process) (SPINSANTI; OSTERMANN, 2013) and surrounded by quality and privacy concerns (ELWOOD; LESZCZYNSKI, 2011; KOUNADI; RESCH; PETUTSCHNIG, 2018), social media data have drawn attention from the research community and emerged as a key issue in the field of disaster management and early warning systems – as pointed out by Steiger, Albuquerque and Zipf (2015), Imran *et al.* (2015), and Landwehr and Carley (2014).

In the particular context of social media analytics for meteorologic issues and flood risk management, however, scant attention has been given to the understanding, characterization, and extraction of spatio-temporal rain patterns. Most of the recent studies have focused on the estimation of flood inundation extent based on the combination of flood patterns extracted from social media data and conventional data (e.g., surveys, ground-gauge data, and imagery data) (HUANG; WANG; LI, 2018; ROSSER; LEIBOVICI; JACKSON, 2017). Nonetheless, rain patterns are fundamental inputs in many rainfall-runoff models for flood modeling and forecasting (e.g., Probability Distributed Model (PDM) (MOORE, 2007)), as well as for early warning systems of extreme weather, such as alerts provided by Civil Defenses and monitoring agencies.

Given the importance of rain patterns and the potential use of social media as a proxy data source of rain phenomenon, this doctoral research investigates the accurate mining of rain patterns from social media activity. Once such patterns have been extracted from social media (geo-social patterns), they can be integrated/assimilated with conventional data sources for practical purposes of monitoring and prediction (RESTREPO-ESTRADA, 2018).

1.2 Challenges and goals

Nonetheless, social media activity is often dispersed in space, irregular in time, and uncertain in content (ZHOU; CHEN, 2014), which means people interact through social media platforms in different places in cities and times during the day and write messages in colloquial language (i.e., with ambiguities and grammar mistakes). On the one hand, social media are filled with over-abundance of data with extensive spatial coverage, sharing a variety of viewpoints/perceptions of the surroundings. On the other hand, such data are discontinuous, uncertain, and noisy.

Several significant challenges should be faced in the extraction of spatio-temporal thematic patterns from social media activity, especially natural phenomenon patterns that are usually continuous in space and take different (spatial) shapes. Although the previous studies demonstrated the potential use of social media as a proxy data source of natural phenomena, some issues have not been thoroughly investigated, such as (i) selection of a representative spatio-temporal scale of analysis, (ii) synchronization of social media activity with the different phases of a given natural phenomenon of interest, (iii) comprehensive understanding behind the production of social media data, and (iv) semantic ambiguities and noises in the information retrieval process. Therefore, they can still be raised as challenges.

Regarding the above-mentioned challenges (i–iv), the goals addressed in this doctoral thesis include:

- (i) **Choice of an appropriate and representative spatio-temporal scale in social media analysis instead of one arbitrary or imposed by conventional data sources.**

Most analyses in social media are based on a spatio-temporal scale chosen with no support of intrinsic and clear criteria that ensure the representativeness of a given natural phenomenon of interest. The scale of analysis is sometimes arbitrarily chosen, or determined/aggregated by matching with conventional data sources (e.g., surveys and physical sensor measurements). Additionally, social media activity operates in space and time differently than instruments that create conventional data sources and, thus, requires a particular way of determining the spatio-temporal scale. This assumption can be made even between different instruments that measure the same phenomenon – for example, precipitation measurements by rainfall gauges and weather radars (each instrument operates in a way and has particular characteristics of coverage and latency). Moreover, the results and conclusions drawn from a social media analysis depend, to a great extent, on the areal and time units chosen, since they face the well-known modified areal unit problem (OPENSHAW, 1984; FOTHERINGHAM, 1989; DARK; BRAM, 2007) and modified temporal unit problem (CHENG; ADEPEJU, 2014). Therefore, systematic approaches for an intrinsic choice of an optimal and representative spatio-temporal scale for mirroring rain patterns from social media activity should be investigated.

(ii) Determination of the temporal validity of spatio-temporal thematic patterns of social media activity regarding a natural phenomenon of interest.

Several studies have investigated the temporal relationship between social media activity and a given phenomenon of interest using only the duration and frequency of time series. In general, they assume messages are generated within the time-span of a given phenomenon. However, such an assumption is valid only for large temporal scales (e.g., one-day scale or greater). Although social media encourages quick status updates and many of them are posted during a given phenomenon of interest, others are posted prior to or after the occurrence of the phenomenon (ANDRADE *et al.*, 2017); i.e., some relevant messages are unsynchronized with the phenomenon of interest (especially on a fine-grained temporal scale). For example, some people forecast rain as “gray clouds in the sky” and “I think it is going to rain”; others refer to past as “It’s stopped raining”. Part of the information from social media activity is likely to be time-invalid for the composition of social signals of natural phenomena. Therefore, an appropriate temporal validation window should be analyzed and established.

(iii) Understanding and characterization of the spatial distribution of thematic social media activity.

Despite the benefits of the use of social media as supplementary data source to conventional data, this data source is heterogeneous in space and time. Several studies have recognized the social media activity is associated with certain groups (e.g., high income and education, and youngers), places (e.g., airports), and periods of the day (e.g., between 8 and 23 hours) (SLOAN, 2017; LI; GOODCHILD; XU, 2013) – which characterizes an uneven spatial distribution of the social media activity. While most of such studies have analyzed the spatial patterns of social media activity only through the perspectives of landmark sites and regions where people live (i.e., by using census data), the way people move around the city has been marginally addressed. Indeed, people post messages on social media platforms when they are moving around the city, and such a dynamic behavior should be taken into account. Besides, thematic patterns from social media may be even more heterogeneous and difficult to be explained only from the perspective of where people live.

(iv) Assessment of the capacity of the keyword-filtering approach to provide reliable spatio-temporal patterns.

Social media is a valuable source of information about human experiences with and within places. People frequently use a wide array of words to report on their experiences, feelings, and observations, as well as their perceptions of a natural phenomenon they have observed or heard about – for example, the words “light rain” and “drizzle” could be interchangeably used to report a fine-rain episode. They are usually employed for the filtering of a large volume of social media messages on a given natural phenomenon of interest. However,

such words may be associated with different or multiple meanings and, thus, filter non-relevant messages for social media analyses. For example, the word “mine” can be a possessive pronoun (“this work is mine”) or a noun (“I work in a mine”). Therefore, a keyword-based filtering approach should be calibrated with keywords that maximize the retrieval of relevant messages.

The challenges and goals above mentioned support the formulation of the major question of this doctoral research:

How can a degree of accuracy and reliability be ensured for the extraction of rain patterns from social media activity?

Certainly, a number of other challenges, including privacy, engagement, and quality concerns, are intrinsically associated with investigations on spatio-temporal thematic patterns obtained from social media data. However, such concerns are outside the scope of this doctoral research.

1.3 Contributions

Regarding goals (i), (ii), (iii), and (iv), this doctoral research aims at supporting the extraction of spatio-temporal thematic patterns from social media activity for reflecting natural phenomena, precisely rain phenomenon, and its contributions include:

- **A systematic method for the selection of analysis scales in social media analytics** ([ANDRADE *et al.*, 2020](#)).

The method, called multicriteria optimization framework, relies on Pareto optimality to assess candidate areal units based on a set of user-defined criteria. A case study investigated heavy rainfall-related tweets and determined the areal units that could optimize spatial autocorrelation patterns through the combined use of indicators of global spatial autocorrelation and variance of local spatial autocorrelation. The results show the optimal areal units (30 km² and 50 km²) provide more consistent spatial patterns than the other areal units evaluated and are likely to produce more reliable analytical results. Nonetheless, the method can be adapted for the selection of a suitable areal unit in different spatial datasets and application cases, including any number of optimization criteria. Its advantage is it optimizes multiple criteria or objective function.

- **A temporal approach for the evaluation of the temporal validity of social media activity related to a given natural phenomenon of interest** ([ANDRADE *et al.*, 2017](#)).

This approach relies on the cross-correlation measure to assess the similarity of social media activity concerning a continuous phenomenon of interest by means of temporal units and their lag-time. The cross-correlation enables the understanding of the temporal relationship between social media activity and the different phases of a phenomenon of interest, and the establishment of a valid temporal window in which messages are relevant for the establishment of social signals. A case study that used rain-related tweets and rainfall measurements provided evidence that rain patterns obtained from rain gauges and Twitter data are not synchronized, but they are linked to a lag-time that ranges from -10 to +10 minutes. This approach can be used for different application cases and domains in social media analytics, since individual data sources can be represented as single time series.

- **A conceptual specification model that characterizes urban areas where the social media activity accurately mirrors rain patterns.**

This conceptual model relies on a linear regression combining demographic, socioeconomic, and human mobility data to characterize places where social media activity accurately mirrors rain patterns. A case study carried out in Sao Paulo, Brazil, revealed human mobility affects the spatial distribution of social media activity. The results indicate the number of trips arriving in the spatial units of analysis, obtained from the Origin-Destination matrix, explains the spatial distribution more than population and income variables. Only the number of trips explains 60% of the variability, which implies human mobility data offer more advantages for the understanding of social media activity than solely demographic and socioeconomic data.

- **A sensitivity analysis for the selection of relevant rain- flood-related keywords** ([ANDRADE *et al.*, 2018a](#)).

A sensitivity analysis carried out assisted the selection of keywords that increase the performance of the keyword-based filtering approach, i.e., keywords that filter more true-positive messages than false-positive ones. Although widely employed to filter social media messages, the approach introduces high noise rates when keywords are not appropriately chosen. Results show some rain- flood-related keywords are more likely to retrieve true-positive messages than others. However, this probability can shift over time and space, since keywords are a product of human experiences with and within places.

Although this doctoral research focuses on rain patterns extracted from social media data, its contributions can be adapted to other phenomena and spatial processes, such as flood analysis with the use of lattices.

1.4 Thesis outline

[Chapter 2](#) describes the multicriteria optimization framework for the selection of an optimal areal unit in social media analytics, as well as a case study conducted for its evaluation. It also introduces the modified areal unit problem and multi-criteria decision analysis, which merit special attention for the understanding of the framework.

[Chapter 3](#) outlines the temporal approach for the assessment of the relationship between social media activity and a given phenomenon of interest, i.e., it establishes the temporal validity of this relationship by means of temporal units and their lag-time, and evaluates the implications of temporal aggregation in this relationship.

[Chapter 4](#) addresses the characterization of the places and introduces aspects of spatial heterogeneity in social media activity where rain patterns extracted from social media data accurately mirror the rain phenomenon. A conceptual model explains the uneven spatial distribution of social media activity through demographic, socioeconomic, and human mobility data.

[Chapter 5](#) examines a case study to measure the capacity of keywords in supporting the keyword-based filtering approach, i.e., from a set of rain- and flood-related keywords, determining those more likely to retrieve true-positive messages.

[Chapter 6](#) provides the conclusions, contributions and limitations of the doctoral thesis, and suggests some future work.

Finally, [Appendix A](#) introduces GIScience and spatial analysis. Nonetheless, for a complete understanding of the fundamental concepts of this doctoral thesis, see [Goodchild \(1992\)](#) and [Goodchild \(2010\)](#) for GIScience discipline, [Singleton, Spielman and Folch \(2018\)](#) and [Batty \(2013\)](#) for urban analytics, and [Anselin \(1988\)](#) and [Haining \(2003\)](#) for spatio-temporal data analysis.

A MULTICRITERIA OPTIMIZATION FRAMEWORK

In establishing the relationship between social media activity and a given real-world spatio-temporal process, the analyst often has to make a decision about which areal unit of aggregation to use. This decision is unavoidably related to the classic and well-studied problems of ecological fallacies and the so-called Modifiable Areal Unit Problem (MAUP) ([OPENSHAW, 1984](#); [FOTHERINGHAM, 1989](#); [DARK](#); [BRAM, 2007](#)). The choice of an areal unit of analysis may be even more complex in social media research than in other areas, since the uneven distribution of social media activity across the urban space is caused by bias in the production practices of social media users and varies in different types of social media platforms ([RZESZEWSKI, 2018](#)). The relationships between the spatio-temporal processes which govern social media activities and spatio-temporal phenomena of interest are poorly understood. The question of which spatial granularity should be used in social media analysis is thus riddled with uncertainty, as the analyst will often be unsure about how to match the areal unit of analysis to the scale of the phenomena being analyzed.

In view of the potentially serious effects of MAUP on social media research and the uncertainty it arouses, it is surprising that investigations into the effects of MAUP on social media analytics have so far received scant attention. However, whilst most previous studies failed to carry out any investigation into the effects of MAUP or justify their areal unit choices, there have recently been a number of studies that either explicitly address or avoid the issues of MAUP ([LEE *et al.*, 2016](#); [JIANG](#); [MIAO, 2015](#)). However, these studies are mostly based on a single criterion (e.g., global measures of spatial association), and fail to adopt a generic approach that takes account of a number of other criteria, such as the need to identify significant local spatial patterns.

This chapter puts forward a systematic approach to support the analyst in investigating the degree of sensitivity to MAUP effects and choosing the most appropriate spatial granularity

for a specific application case study. It establishes a multicriteria optimization framework to assist in the selection of the areal unit in social media analysis, which is based on the definition of a number of criteria (e.g. global and local indicators of spatial association) and the application of the Pareto optimality method. Pareto optimality has been widely used to assess a number of alternative solutions in problems that involve multiple criteria, and where a solution that is regarded as ‘optimal’ for one criterion may not be for another. Multiple conflicting criteria can thus be evaluated to answer questions such as the following:

- How can we ensure that the ‘optimal’ spatial unit chosen suitably characterizes or represents the spatial process in accordance with a number of given criteria?
- What is the ‘optimal’ spatial unit that should be used when there are multiple and conflicting criteria?

This multicriteria framework is, thus, applied to investigate the effects of different areal units on the analysis of heavy rainfall patterns by means of Twitter data in the city of Sao Paulo, Brazil. The remainder of the chapter is structured as follows: [Section 2.1](#) provides an overview of the literature on the effects of MAUP on social media analyses; [Section 2.2](#) outlines the multicriteria optimization framework for the selection of an appropriate spatial unit in social media analysis; [Section 2.3](#) describes a case study on the use of social media as a proxy for heavy rainfall patterns; [Section 2.4](#) and [Section 2.5](#) report the main results and conduct a discussion on the findings; finally, [Section 2.6](#) presents the final remarks.

2.1 MAUP in social media analysis

The spatial approach to social media analysis often involves aggregating messages to a study area which has been partitioned into areal units that vary in size, from square meters to square kilometers, and shape, such as regular and irregular polygons. This kind of spatial arrangement for data aggregation is sensitive to the scale and zoning effects of MAUP, which can yield different spatial patterns and statistical results owing to uncertainty about the number (scale effect) and shape (zoning effect) of the areal units ([OPENSHAW, 1977](#); [OPENSHAW, 1978](#); [OPENSHAW, 1984](#); [FOTHERINGHAM, 1989](#); [DARK; BRAM, 2007](#)).

[Figure 2](#) shows the influence of the scale and zoning effects of MAUP for spatial data aggregation in a given study area. Clearly, the density patterns reported for any one particular areal unit ([Figure 2](#) (c) and (d)) could be misleading if taken as representative of the sample of spatial data bounded to the study area ([Figure 2](#) (a)). For example, the densities of spatial data across the vertical arrangement in [Figure 2](#) (c) varies considerably when compared with the constant densities across the horizontal arrangement. A similar behavior can be seen in [Figure 2](#) (d), where the low density appears on horizontally opposite sides when the two spatial

Figure 2 – Scale and zoning effects for spatial data aggregation. (a) corresponds to a sample of spatial data at an individual level bounded to a study area, (b) is the spatial count data using a reference lattice design, (c) and (d) illustrate the scale and zoning effects, respectively.

Source: Adapted from [Lee et al. \(2016\)](#).

arrangements are looked at together. It is widely recognized that, different conclusions can be drawn about the underlying statistical relationships depending on the choice of an areal unit of analysis ([FOTHERINGHAM, 1989](#)). As a result, if no systematic criteria are used for the assessment the effects of MAUP and for the choice of an areal unit of analysis, this may lead to the data being aggregated in a biased and mistaken zoning system.

Although it is widely regarded as a problem that is inherent to spatial analysis, the literature provides some possible strategies for dealing with MAUP ([DARK; BRAM, 2007](#); [FOTHERINGHAM, 1989](#)). These include the following:

- (i) the derivation of an ‘optimal’ zoning system where a hypothesis concerning the expected results can be attained ([OPENSHAW; RAO, 1995](#); [OPENSHAW, 1977](#));
- (ii) the identification of basic entities and primitive areal units as a means of avoiding the use of data aggregation ([JIANG; BRANDT, 2016](#));
- (iii) the development of new methods that lay greater emphasis on visualisation than statistical analysis ([TOBLER, 1989](#));
- (iv) the emphasis of spatial analysis on the rates of change ([FOTHERINGHAM, 1989](#); [POORTHUIS, 2018](#)); and

- (v) the sensitivity analysis that examines the effects of MAUP by reporting the results for different areal units (FOTHERINGHAM; WONG, 1991).

Although previous approaches have proved effective in understanding and addressing MAUP, they tended to deal with special cases of a general problem and should be applied with some caution – depending on the project and type of analysis (DARK; BRAM, 2007).

MAUP is often ignored in social media analytics and empirical studies involving the analysis of areal data rarely mention possible scale and zoning effects. This is especially true in urban analytics that use social media data around a topic to mirror real-world spatio-temporal phenomena – for some examples, see Restrepo-Estrada *et al.* (2018), Arthur *et al.* (2018), Tenkanen *et al.* (2017) and Longley and Adnan (2016). However, there have recently been a number of studies that clearly address the question of MAUP. For example, Jiang and Miao (2015) delineated urban boundaries of cities by means the topology of social media activity. They used the heterogeneity of the hierarchical agglomerations of social media activity to determine the urban structure, which may mitigate the statistical bias of MAUP. However, this work does not make a systematic assessment of MAUP effects to provide evidence of improvements achieved by its selection strategy for the spatial unit of analysis. In contrast, Lee *et al.* (2016) assessed the scale effect of MAUP through the rate of change of an indicator of global spatial association (Global Moran's I) using regular grid lattices with different areal unit sizes. Analogously to a previous work in the segmentation of high resolution remotely sensed images (Meng *et al.*, 2014), Lee *et al.* (2016) proposed to select the areal unit of analysis based on the lattice layout that yields the higher Global Moran's I coefficient.

Although the use of global indicators of spatial association for a sensitivity analysis has proved to be a useful way of investigating MAUP effects, this method only considers the overall clustering patterns of georeferenced social media data, whilst the spatial variance or structural instability of local patterns has been neglected. Global Moran's I coefficient alone may not be enough to diagnose the spatial heterogeneity of social media activity, particularly in study areas partitioned into a large number of areal units of analysis. One of the reasons for this is that global patterns of spatial association usually assume spatial homogeneity (ANSELIN, 1995) and social media activity is often dispersed in space, irregular in time, and uncertain in content. Moreover, local spatial patterns may be of particular relevance in urban analytics due to the intra-urban inequalities that influence user-data generation. As a result, social media activity is often associated with a low/medium spatial dependence (i.e., a degree of spatial association) and a high level of spatial heterogeneity. Thus, the investigation of the effects of MAUP should take account of other indicators, such as the spatial heterogeneity of the process in a study area.

An example which takes the structural instability of the local patterns into account is given in Figure 3. Each lattice (Figure 3 (a)) is related to a global indicator of spatial association and the standard deviation (i.e., spatial variance) of the local indicators of spatial association.

These spatial association statistics were calculated by means Global Moran's I coefficient and its local version. According to [Anselin \(1995\)](#), local indicators of spatial association (LISA) are spatial decomposition statistics of global indicator of association that enable the identification of spatial outliers and make an assessment of the overall structural (in)stability, which is useful for an analysis of spatial heterogeneity. If the underlying process is stable throughout the lattice, the local indicators are expected to show a constant statistical behavior at the areal unit chosen.

As shown in [Figure 3 \(b\)](#), spatial variance changes considerably and a high Global Moran's I does not necessarily yield a low standard deviation of Local Moran's I , i.e., the choice of the most suitable lattice should be bi-dimensional in terms of global and local statistics. Hence, the spatial heterogeneity analysis is of value for assessing the extent to which a global indicator is representative of the local association ([ANSELIN, 1995](#)) and measuring the conflict between spatial stability and the global indicator of association. This kind of trade-off can occur in any spatial data analysis, including urban social media analytics in different periods and areal units.

Figure 3 – Trade-off between Global Moran's I and the overall degree of structural (in)stability (standard deviation) of Local Moran's I . It should be noted the difference when compared the trends of the standard deviation of Local Moran's I and Global Moran's I across the lattices. Standard deviation of the Local Moran's I was normalized by scaling between the minimum and maximum values of the Global Moran's I . Both statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.

Source: Elaborated by the author.

In summary, the determination of an optimal areal unit for spatial analysis of social media data is a complex task owing to the MAUP effects, differences in the fields of application, and uncertainties and conflicts arising from the different potential spatial indicators to be used. Since a global (or singular) optimal areal unit cannot be determined, the approach I adopt to address this problem is to enable the assessment of various different areal units by multiple indicators in order to subsequently support the selection of an optimal areal unit, depending on the application and determination of the spatial analysts.

2.2 Multicriteria optimization framework

Figure 4 shows the multicriteria optimization framework established by a Multi-Criteria Decision Analysis (MCDA) for the choice of an optimal areal unit in spatial data analysis, which focuses on studies involving the social mirroring of real-world phenomena derived from social media activity. MCDA is a discipline that provides a systematic and generalized way of dealing with decision problems, by assisting decision-makers to choose an appropriate and satisfactory solution from a finite set of candidate or alternative solutions (GRECO; EHRGOTT; FIGUEIRA, 2016; XU, 2012). According to Xu (2012), MCDA ‘refers to making decisions in the presence of multiple, usually conflicting, criteria’. As argued in Section 2.1, the choice of an areal unit in social media analysis related to real-world phenomena is closely linked to the evaluation of conflicting indicators or criteria. Hereinafter, the words indicators and criteria will be used interchangeably in the context of MCDA.

Figure 4 – Methodological multicriteria optimization framework for the selection of an optimal areal unit in a spatial data analysis.

Source: Elaborated by the author.

Modelling of candidate areal units

An MCDA problem can be modeled by a bidimensional decision matrix in which each element (cell) represents the outcome of a measure against a criterion (column) and corresponds to a particular decision – also referred to as a candidate solution (line). The number of criteria and candidate solutions is unlimited; however, both can be reduced if knowledge is drawn from the project topic and type of analysis. In problems concerning urban analytics, this means choosing a range of areal units that are geographically meaningful, and spatial statistics that make sense to

the problem/analysis in hand – the analyst should reduce the search space of the candidate areal units within the multicriteria optimization framework.

In general, the establishment of criteria depends on the problem, and no set rule is followed. I assessed the areal units in social media on the basis of two criteria, namely Global Moran's I and the coefficient of variation of Local Moran's I . The former relies on the spatial aspects of a global social media activity (i.e., the average of the overall spatial patterns), whereas the latter measures the overall instability through local inequalities (i.e., the variance of the local spatial patterns). These spatial statistics can be calculated by means of different schemes of spatial contiguity and spatial weight matrices, but I computed them for a row-standardized spatial weights matrix based on first-order rook contiguity (i.e., adjacent neighbors) – since the first-order rook makes sense for the case study (i.e., the mapping of a continuous phenomena) and the results remained stable across different schemes of spatial weights matrix. The coefficient of variation was used to summarize the Local Moran's I , rather than the standard deviation, since it allows direct quantitative comparisons to be made between different probability distributions, i.e., comparisons between spatial variances of Local Moran's I across different areal units.

Evaluation of the candidate areal units

Although the MCDA methods share similar modelling procedures (i.e., stages in organization and decision matrix construction), they synthesize and optimize the criteria, and calculate the decision matrix differently (GRECO; EHRGOTT; FIGUEIRA, 2016). Hence, selecting a particular MCDA method depends on the characteristics of a given problem. Collette and Siarry (2004) and Greco, Ehrgott and Figueira (2016) provide a review of the well-established and recently emerging fields, theories and methods within MCDA, which assist the readers in linking problems to methods. I have selected the Pareto optimality algorithm available in rPref package (ROOCKS, 2016), which is a dominance-based method. In general, it sorts the candidate solutions into Pareto frontiers based on all the trade-offs of the criteria and leaves the selection of a preferred candidate solution free for the decision-maker. Frontiers are cutting points that group the candidate solutions into ordered classes that range from the best (first frontier) to worst (last frontier). All the candidate solutions that fall into the same frontier are considered to be interchangeable. The so-called Pareto-optimal solutions are those that fall into the first frontier, which are assumed by the method to be the most suitable solutions.

Pareto optimality method

Let X be a set of user-defined areal units with different levels of aggregation. Each spatial granularity of aggregation $x \in X$ is characterized by different criteria that will be optimized by a set of objective functions. A vector containing m objective functions ϕ_m can be represented by

$$\Phi(x) = [\phi_1(x), \phi_2(x), \dots, \phi_m(x)] \in \mathbb{R}^m \quad (2.1)$$

A Pareto-optimal solution only contains areal units that are not Pareto-dominated by any other areal unit. More formally, but still in general terms, an areal unit $x_i \in X$ dominates another $x_j \in X$ when it has satisfied the following two constraints:

$$(i) \quad \forall \phi \in \Phi : \phi(x_i) \preceq \phi(x_j), \text{ and}$$

$$(ii) \quad \exists \phi \in \Phi : \phi(x_i) \prec \phi(x_j)$$

where $\prec$ and $\preceq$ correspond to the ‘general better’ and ‘better or equal’ relations, depending on whether the objective function refers to maximization or minimization. All the Pareto-optimal areal units form the first Pareto Frontier and if two or more areal units fall into it, additional human expertise is required for the selection of a proper areal unit. As mentioned above, all the areal units in the first Pareto Frontier are considered to be equally ‘good’. The other frontiers are calculated in the same way, although the areal units of the previous frontiers are removed (e.g., the second frontier is calculated by removing the areal units of the first frontier, the third frontier is calculated by removing the areal units of the first and second frontiers, and so on).

Sensitivity analysis of the optimal areal units

Once the first frontier has been obtained, the robustness of its solutions must be evaluated. Within the context of MCDA, a sensitivity analysis is a common approach for investigating the statistical robustness of Pareto-optimal solutions (FONSECA; FONSECA; HALL, 2001). A practical way of carrying this out, it is to check the stability of the outcomes obtained from multiple runs of the Pareto optimality algorithm. Random resampling or disturbances from the original data should be introduced to give an idea of how stable (i.e., robust) the Pareto-optimal solutions in each run tend to be. In line with this approach, I applied a bootstrap method since this has been recognized as an asymptotic resampling approach in different contexts (EFRON, 1979).

Figure 5 shows the bootstrap resampling strategy used to generate samples of spatial data; it uses blocks of data to partially ‘retain’ the original spatial properties. Each block corresponds to an event/occurrence of a particular phenomenon and shares a set of data (e.g., geotagged social media messages related to the given rain phenomenon on a rainy day). In this work, an event is understood to be a measurement/observation of an observed phenomenon within a study area (e.g., daily or hourly observations of rainfall in a city). Hence, a resampling of events (i.e., sets of social media data grouped into time units) were generated to perform the sensitivity analysis.

Figure 5 – Bootstrap resampling strategy with set of spatial data grouped into regular time units called events. • corresponds to spatial data at an individual level spread over a study area (e.g., geotagged social media messages across the city of Sao Paulo).

Source: Elaborated by the author.

2.3 Application of the multicriteria optimization framework

Case study in the context of heavy rain in Sao Paulo city, Brazil

The multicriteria optimization framework was employed for the selection of the optimal areal unit in a social media analysis within the context of heavy rainfall patterns in Sao Paulo city, Brazil. The city was chosen because heavy rain events cause flash floods, inundations and flooding, mainly due to the rivers overflowing and the poor drainage system of the urban pavements. It should also be noted that, Sao Paulo has a vast number of Twitter users and an estimated population of approximately 12 million people, which has made it the most density populated city in Brazil (IBGE, 2010).

The entire surface area of the city was partitioned into hexagonal areal units of 5 km², ranging from 10 to 100 km² every 10 km², and 200 km². Each spatial unit aggregated rainfall data and rain-related Twitter messages over the period of 1 year, from November 2016 to November 2017. This range of areal units allowed us to determine the trade-off between the Global Moran's *I* and the coefficient of variation of Local Moran's *I* for a specific application case, while the hexagonal areal units reduced the visual field bias when compared with the square units (CARR; OLSEN; WHITE, 1992). Moreover, hexagonal areal units also favor the construction of the spatial contiguity matrix over square areal units by the number of natural boundaries.

Description of the data

Twitter data

I used the Twitter Streaming API to fetch public geotagged tweets that fell within Sao Paulo city. Although the methods employed by Twitter Streaming API for sampling data are unknown, they return a large enough set of geotagged tweets from the Twitter population (MORSTATTER *et al.*, 2013). 2,073,219 million geotagged tweets were sampled within the city during the entire period of analysis. Although I examined a large and dense dataset, the geotagged tweets related to the rain phenomenon represented a small fraction (5,996 / 0.29%) of the total number of geotagged tweets. A low percentage of phenomenon-related geotagged tweets was also observed in other studies on crises and natural disasters (XIAO; HUANG; WU, 2015; HUANG; XIAO, 2015; ALBUQUERQUE *et al.*, 2015).

Five meaningful rain-related keywords obtained from Andrade *et al.* (2018a) – ‘chuva’, ‘chove’, ‘chuvoso’, ‘chuvosa’ and ‘chuvarada’ (in Brazilian Portuguese) – were employed and any geotagged tweets containing at least one of them were filtered (i.e., the filtered messages were ‘labeled’ as related to the rain phenomenon). The authors showed that these keywords are less sensitive to time and space than others and thus have the potential to create a filter that produces more signal than noise, i.e., they are almost invariant across space and time (when the study area is a city) and filter more true-positive (signal) than true-negative (noise) Twitter messages. True-negative tweets are referred to as those that contain at least one keyword and where the text content is not linked to the phenomenon of rain. An example of a true-negative tweet is ‘bolinho de chuva’ (little rain cookie), a typical Brazilian doughnut.

I built a rainfall signal on the basis of the filtered geotagged tweets by means of the ‘odds ratio measure’ of the frequency of geotagged tweets (Equation 2.2) on a one-day scale.

$$OR = \frac{p_i/p}{r_i/r} \quad (2.2)$$

where p_i is the number of rain-related geotagged tweets in a spatial unit i , p is the total number of rain-related geotagged tweets, r_i is the number of geotagged tweets in a spatial unit i , and r is the grand total of geotagged tweets. This kind of measure is based on the Location Quotient technique and takes into account the sampling Twitter data to offset the over-representation of the space units, which reduces the effects of a dissimilar spatial distribution of Twitter activity (POORTHUIS *et al.*, 2014). In addition, a higher odds ratio measure leads to a better representation of rainfall signals with regard to the Twitter activity within the spatial unit of observation.

Rainfall maps from the weather radar

The Sao Paulo weather radar of the Department of Water and Power (DAEE) and the Hydraulics Technology Foundation Center (FCTH) of the Polytechnic School of the University of Sao Paulo diagnosed the rain maps every 5 min. This device is a Dual Polarization Doppler S-Band weather radar located approximately 60-70 km from Sao Paulo city which continuously estimates the statistics of rainfall rates at a 250 meter spatial resolution from an azimuthal width of 1 degree (SELEX, 2015).

Instead of using the polar rain yields, I relied on a radar product known as Surface Rain Intensity (SRI) projected in Cartesian coordinates of a 500x500 meter scale (SELEX, 2015). This enabled us to create new maps at 10 min. time intervals to match the representation of rainfall signals used in the Twitter activity. The mean, minimum and maximum values of each time interval were stored in each spatial observation unit. Spatial and linear interpolation techniques were employed to overcome the problem of missing data for measurements within the same day. Despite errors in the weather radar rainfall estimates, such as overestimation of some observation points (BATTAN, 1973), the radar measured the rainfall rate accurately.

On the basis of the rainfall threshold of the U.S. Geological Survey (USGS), I selected days when the daily record of rainfall was, at least, higher than 10 mm per hour, i.e., I took note of any rain event equal to, or greater than, a heavy shower. This threshold is close to the one generally used by Brazilian meteorological centers for heavy rain, such as the Brazilian National Center of Monitoring and Early Warning of Natural Disaster (CEMADEN). Figure 6 shows the daily increase the rainfall and frequency of rain-related tweets of a cross-sectional data from 7 November 2016 to 26 April 2017.

2.4 Results

Optimal areal units

Figure 7 shows Global Moran's I coefficient and the coefficient of variation of Local Moran's I for the areal units. Only some of the areal units show an improvement in the criteria when compared with the adjacent areal units, i.e., higher Global Moran's I and lower coefficient of variation of Local Moran's I . The other areal units either increase or decrease both criteria. For example, from 20 km² to 30 km² both criteria improved, i.e., Global Moran's I coefficient increased and the coefficient of variation of Local Moran's I decreased. This means that the areal unit of 30 km² is linked to a higher pattern of spatial association and lower spatial heterogeneity than the areal unit of 20 km², i.e., the former provides more consistent spatial patterns and is thus likely to reflect more reliable analytical results. A similar improvement was achieved by the other areal units, such as from 80 km² to 90 km² and from 90 km² to 100 km² (Figure 7).

Figure 6 – Cross-section data of daily rainfall and frequency of rain-related geotagged tweets from 7 November 2016 to 26 April 2017, Sao Paulo, Brazil.

Source: Research data.

Figure 7 – Trade-off between the global indicator of spatial association (Global Moran's I) and the overall degree of structural (in)stability (coefficient of variation of Local Moran's I normalized by scaling between the minimum and maximum values of the Global Moran's I coefficients). Both global and local spatial statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.

Source: Research data.

In contrast, the areal units of 30 km² and 50 km² appear to achieve the best results in visual terms, although the criteria are in conflict with each other. While Global Moran's I coefficient is higher for the areal unit of 50 km², the coefficient of variation of Local Moran's I

is lower for the areal unit of 30 km^2 , i.e., if the criteria had been analyzed separately, they would have led to different results. The areal unit of 50 km^2 is better in terms of overall spatial patterns, but the one of 30 km^2 is better in terms of the structural stability of the local patterns. A similar conflict also appears between the areal units of 100 km^2 and 200 km^2 . In view of this, more conflicts might occur, depending on the range, size and shape of the user-defined areal units.

According to the results of the multicriteria optimization framework in Figure 8, Pareto-optimal solutions are the areal units of 50 km^2 and 30 km^2 , since both areal units fall within the first Pareto frontier and are thus considered to be equally good (see Section 2.2). The Pareto front line visualization (Figure 8) is a ‘stair-shaped line marking the border of the dominance area’ of those tuples (solutions) (ROOCKS, 2016). For example, the areal units of 50 km^2 and 30 km^2 dominate the other ones because their criteria are better, i.e., they are combined with a higher Global Moran’s I and lower coefficient of variation of Local Moran’s I , respectively. However, they are mutually dominated by one criterion, i.e., the areal unit of 50 km^2 dominates 30 km^2 in terms of Global Moran’s I and the opposite is the case for the coefficient of variation of Local Moran’s I (see Table 1). Such a mutual domination is illustrated in Figure 8 by a link (line) between the scales.

Figure 8 – Pareto frontier and trade-off between Global Moran’s I and the coefficient of variation of Local Moran’s I (overall degree of structural (in)stability). Both statistics were computed for a row-standardized spatial weights matrix based on first-order rook contiguity.

Source: Research data.

Sensitivity analysis of the optimal areal units

Figure 9 shows the Pareto-optimal areal units (i.e., those within the first frontier) of a thousand runs of the multicriteria optimization framework performed by the bootstrap method which is used as a resampling approach – for details of the bootstrap resampling strategy, see

Table 1 – Criteria values and the corresponding Pareto frontier for each areal unit. In the case of Global Moran's I criterion, the p -value corresponds to the null hypothesis that the social media pattern is randomly distributed among the spatial units.

Frontier	Areal unit	Global Moran's I (p -value)	$\Delta_{\text{Coef. Var.}}$ of Local Moran's I
#1	50 km ²	0.40 (0.001)	± 1.93
#1	30 km ²	0.36 (0.001)	$\pm \mathbf{1.65}$
#2	40 km ²	0.32 (0.001)	± 2.79
#2	200 km ²	0.31 (0.028)	± 2.17
#2	100 km ²	0.28 (0.011)	± 2.02
#3	90 km ²	0.26 (0.017)	± 2.56
#4	70 km ²	0.19 (0.021)	± 2.91
#5	80 km ²	0.17 (0.048)	± 4.17
#5	60 km ²	0.06 (0.102)	± 3.22
#6	20 km ²	0.03 (0.170)	± 7.81
#7	5 km ²	0.02 (0.120)	± 21.82
#8	10 km ²	0.01 (0.268)	± 47.19

Source: Research data.

Section 2.2. The areal units of 50 km² and 30km² were significant and corresponded to 32.5% and 28.2% frequencies, respectively (Table 2); they covered 60.7% of the outcomes. Thus, they can be regarded as robust Pareto-optimal solutions within the scope and limitations of the case study.

Figure 9 – Robustness of the Pareto-optimal areal units using the bootstrap method with 1,000 replications – for details of the bootstrap resampling strategy, see Section 2.2.

Source: Research data.

Table 2 – Pareto-optimal areal units and corresponding frequencies after the bootstrap method with 1,000 replications.

Areal unit	Frequency ^(a)	(%)
50 km ²	608	32.5
30 km ²	529	28.2
100 km ²	272	14.5
200 km ²	222	11.9
40 km ²	112	6.0
90 km ²	84	4.5
60 km ²	18	1.0
80 km ²	17	0.9
70 km ²	9	0.5

Note – ^(a) The sum of the frequencies is greater than 1,000 because the Pareto-optimal frontier may contain one or more areal units.

Source: Research data.

Implications of the choice of an arbitrary areal unit

Figure 10 shows the spatial patterns of the heavy rainfall signal from geotagged tweets on the maps of 10 km², 20 km², 30 km², 50 km², 60 km² and 70 km². If an arbitrary areal unit is chosen, such as 20 km² or 60 km², different and discordant spatial patterns are obtained when compared with the Pareto-optimal areal units, i.e., the patterns of the areal units of 30 km² and 50 km² are more similar than the others. In practice, this affects the conclusions and may lead to misunderstandings and mistakes by decision-makers when applying the strategy (i.e., an arbitrary choice) to urban planning (e.g., for the use of social media as a supplementary data source for heavy rainfall and flood risk management).

Figure 10 – Comparison of spatial patterns of Pareto-optimal areal units (middle) and four arbitrary areal units (extremes). The patterns correspond to the ‘odds ratio measure’ of the frequency of geotagged tweets (POORTHUIS *et al.*, 2014). ℓ corresponds to the side length of hexagonal lattices.

Source: Elaborated by the author.

2.5 Discussion

The results of the spatial analysis of geotagged social media (e.g., those related to social mirroring of heavy rain events in the application example) are strongly dependent on the areal unit of analysis. In light of this, a large number of possible areal units must be assessed to determine which is most suitable, and computational techniques can automate their choice by evaluating one or more criteria. This is related to the pioneering work of Openshaw (1977), which minimized a loss function of a linear regression model for the selection of an optimal zonal system in census studies. More recently, in the context of social media, Lee *et al.* (2016) maximized Global Moran's I as a criterion for the selection of an optimal areal unit. However, a limitation of both of these studies is that they rest on the assumption that the most suitable areal unit can be determined on the basis of a single optimization criterion (or loss function). As the example in this chapter shows, a number of different factors may determine the choice of an areal unit and this can lead to divergences, and thus make the assessment a laborious process for the spatial analyst.

In this chapter, I argue that the choice of the most appropriate areal unit for social media analysis (as a data source to study real-world processes) should be based on data-driven approaches that require a multiple-criteria analysis (Section 2.1). The analysis carried out has shown the conflicts that can occur if multiple criteria are employed and I put forward a way of dealing with them that involves establishing a multicriteria optimization framework which relies on the Pareto optimality method (Section 2.2). This framework provides an innovative method for selecting a more appropriate areal unit of aggregation in social media based on multiple criteria, instead of a choice made through an awareness of the analyst or the evaluation of a single criterion (which were the approaches commonly used in previous work). This systematic and novel approach for the selection of an areal unit can be adapted to include any number of optimization criteria that may be appropriate for different spatial datasets and applications. The example of an application for the analysis of social media messages about heavy rain events, shows that the resulting optimal areal units of 30 km² and 50 km² provide more consistent spatial patterns than the other areal units.

Nevertheless, a number of caveats and suggested extensions of this study should be discussed. There exists a relation among the criteria chosen and the results. An 'optimal' areal unit for one particular criterion may not be suitable for another, e.g., the results vary when Global Moran's I and the coefficient of spatial variation of Local Moran's I are optimized individually (Figure 8 and Table 1). For this reason, the knowledge on the project topic and type of analysis should be used for both definitions of the range of user-defined areal units and for defining the criteria that must be evaluated (Figure 4). In this work, Global Moran's I and the coefficient of variation of Local Moran's I have been defined because the evaluation of both criteria together reveals the continuous spatial processes more suitably than the use of a single criterion. Spatial patterns of heavy rain usually involve a high degree of spatial dependence and low level of spatial

heterogeneity and are thus expected to be continuous in space.

Regarding the calculation of spatial statistics of Global and Local Moran's I , a doubt arises on how I should choose the spatial contiguity and spatial weights matrices. In the case study, a row-standardized spatial weight matrix based on first-order rook contiguity was chosen, that was based on the results that remained stable across different spatial weight matrices. However, for a generalized and unbiased use of these spatial statistics, future studies should investigate to what extent the relationship and degree of intensity influences the underlying spatial units (ANSELIN, 1995).

Furthermore, the case study is bounded to the spatial dimension and only addressed the scale effect of MAUP by means of a fixed temporal unit, i.e., the variation in statistical results between different levels of aggregation (i.e., areal unit sizes) in a one-day scale. Although the multicriteria optimization framework can be applied to assess different combinations of areal units (i.e., the scale and zoning effects, as described in Section 2.1), further investigation is required to address the temporal dimension. There are studies that have recently demonstrated the temporal effects of aggregation, segmentation, and boundary in a time series as a means of understanding and analyzing real-world spatio-temporal processes (CHENG; ADEPEJU, 2014). Hence, these temporal effects should be taken into account in a future extension of the framework and case studies.

It should also be noted that when a spatial analysis uses regular lattices, it is often driven by data aggregation approaches and may 'reshape' the topology of the social media messages – depending on the size and shape of the areal unit of aggregation. The topology of social relationships can be useful for identifying landmark sites (e.g., airports, major public parks, and green spaces) and could be better reflected by using bottom-up approaches such as the concept of 'natural cities' (JIANG, 2015; JIANG; BRANDT, 2016), which relies on the head/tail breaks algorithm to delineate irregular areal units and hotspots in a natural setting. Instead, I am interested in the social mirroring of a real-world spatio-temporal process from social media data (by assuming independence of social media messages).

Certainly, a number of other factors are likely to exert influence on the spatial patterns of the Twitter activity, including the topography of the urban space. However, previous work (ANDRADE *et al.*, 2017; RESTREPO-ESTRADA *et al.*, 2018) provides quantitative evidence that rainfall patterns from meteorological data (e.g., rainfall gauges) are significantly associated with rainfall-related Twitter activity in the city of Sao Paulo. This enables us to suggest that the patterns identified in the current application case for the method proposed can be useful for the analysis of the spatial distribution of rainfall within the urban space, even if I recognize the potential for spatial biases introduced by other factors or physical processes (such as topography), which should be further investigated in future work.

Furthermore, the spatial heterogeneity of Twitter activity within the urban space could be associated with the existence of multiple spatial scales in the data (WESTERHOLT; RESCH;

ZIPF, 2015). Addressing this issue goes beyond the scope of the current work, but one potential extension of our approach in this direction could involve assessing the spatial heterogeneity of the Twitter patterns to determine more homogeneous subareas, in which our approach could then be employed for selecting a most suitable local scale. However, I contend that a global optimization approach still makes sense in our application case, since it is used in hydrology for the analysis of rainfall (BLÖSCHL; SIVAPALAN, 1995) and it can thus be useful in practical scenarios of precipitation regimes and flood monitoring.

With regard to the robustness of Pareto-optimal solutions, the areal units of 50 km² and 30 km² covered 60.7% of the bootstrapping method outcomes (Figure 9), and only reflect the findings of a particular application in a case study. As a consequence, the results can change in function of different data, periods and case studies.

Last, but not least, some MCDA modelling methods cannot model the uncertainties that arise from more complex decision-making problems which are often of a quantitative and qualitative nature and extend the application of average values to represent the criteria for an assessment as a distribution (XU, 2012), e.g., the stratification of the average values into classes. In this case, belief structures, such as Evidential Reasoning (ER) approach (XU, 2012), might be employed to model and represent the problem as an MCDA.

2.6 Final remarks

The ‘optimal’ areal unit in social media analysis is traditionally established in different ways through the view of the analyst and the field of application. Frequently, a preferred areal unit is chosen with no clear criteria or justification for ensuring its representativeness regarding the phenomenon under study. In this case, the areal unit chosen may raise methodological issues, such as the MAUP effects. In this chapter, I argued that the ability to select the most appropriate areal unit in spatial analysis should be based on data-driven approaches that rely on a multiple-criteria analysis. A multicriteria optimization framework was established for the selection of an optimal areal unit in social media analysis driven by the Pareto optimality method.

I expect that the choice of an areal unit by adopting a multicriteria approach can provide much more realistic patterns than a unidimensional approach. Indeed, solutions of real-world problems are naturally driven by multiple perspectives (i.e., criteria), which make this multicriteria optimization framework a valuable tool for selecting an optimal areal unit like those described in the application example in social media (Section 2.3). The potential benefits of using it were examined and discussed in detail in Section 2.1, Section 2.4 and Section 2.5. However, the degree of uncertainty that an optimal areal unit can be found increases with the number of criteria and more in-depth expertise is necessary before a decision can be made on the choice of the ideal areal unit.

Future studies should aim at evaluating this framework by exploring other case study

scenarios (e.g., other cities and datasets with more control than social media data) and criteria (e.g., economic and demographic factors) so that generalized results can be obtained. Furthermore, the MCDA discipline encompasses a wide range of algorithms, and other ones can be considered for the handling of multiple and usually conflicting criteria. Finally, forecasting spatial models can be designed on the basis of the optimal areal units provided by the multicriteria optimization framework to further explore the practical use of the approach proposed.

A TEMPORAL VALIDITY APPROACH

In addition to determining the most suitable spatio-temporal scale that establishes the relationship between social media activity and a given natural phenomenon of interest (see [Chapter 2](#)), the analyst also needs to determine the temporal validity of this relationship. This temporal validity is important for establishing the correlation between social media activity and the different phases of the natural phenomenon of interest, especially when we are dealing with natural phenomena that occur quickly and in short time-spans such as torrential rain.

The problem of temporal validity can be understood in terms of the existing timing differences between two or more time series that measure the same natural phenomenon. To put it differently, we know that social media activity can occur at the beginning, middle, or end of a given natural phenomenon – or at all phases of disaster management ([IMRAN *et al.*, 2015](#); [CARLEY *et al.*, 2016](#)) –, but “how can the appropriate time for using the social media data be established?” This question can be answered through a temporal validity of data using a lag-time of the time series. The temporal validity allows us to take note of the existing correlation between social media activity in different phases of a given natural phenomenon of interest, and thus is useful in the field of disaster management. Hence, the time series analysis must include the lag-time variable to ensure it is more effective. In addition, understanding and identifying the lag-time can serve to support the extraction of key information and improve the use of social media data in practical solutions, such as monitoring and early warning systems (e.g., by determining the length parameter in the sliding window approaches).

This chapter adopts a temporal approach to support the analyst in determining the temporal validity between social media activity and a given natural phenomenon of interest. This approach is based on the assumption that the temporal relationship between social media activity and the natural phenomenon of interest can be explained by duration, frequency and the lag-time of the time series, rather than only the duration and frequency of the time series as has been done in social media analytics. Regarding evaluation, a case study was conducted in the city of Sao Paulo, Brazil, using rainfall data and Twitter data related to the rain phenomenon.

The remainder of the chapter is structured as follows: [Section 3.1](#) introduces the temporal validity problem; [Section 3.2](#) gives a detailed account of the case study used for evaluating the temporal approach; [Section 3.3](#) examines the methodology employed; [Section 3.4](#) and [Section 3.5](#) report the main results and draws some conclusions on the findings; finally, [Section 3.6](#) presents the final remarks.

3.1 The temporal validity problem

[Figure 11](#) and [Figure 12](#) depict, to a limited extent, the increase in the rainfall measurements and frequency of rain-related geotagged tweets in two periods of January 2016, in Sao Paulo, Brazil, during the peak of the rainy season. As can be seen, there is a similarity (based on the peaks and valleys) between the rainfall time series and rain-related geotagged tweets time series. However, there is no exact correspondence between both time series. In ordinary circumstances, social media users often report messages related to rain episodes. That is, they post texts, photos and videos, for example, of gray clouds, drizzle, rainfall, and floods, and on rare occasions, give a forecast of rain. These posts appear at different times, i.e., the temporal scale and frequency of rain-related social media messages are not known. They may occur before or after the rain, whilst other occur during a period of rainfall ([Table 3](#)). Moreover, the social media users can forward older information instead of new information ([EARLE; BOWDEN; GUY, 2012](#)).

On the other hand, there is a well-known temporal scale for rainfall data ([BLÖSCHL; SIVAPALAN, 1995](#)). The information usually comes from physical sensors (e.g., rain gauges and weather radars) installed in urban areas. In Brazilian urban areas, for instance, a widely-used temporal scale for rain gauges and weather radars is 10 min. ([Figure 11](#) and [Figure 12](#)). This means that it is a challenge to explore the relationship between the rainfall data and rain-related social media activity for applications of monitoring and early warning.

Table 3 – Examples of geotagged tweets related to the rain phenomenon and classified per time-span (i.e., before, after and during the rain phenomenon).

Datetime	Rain-related tweets	Translation	Time
2016-01-12 06:11:39	“Escuro e quente, a tarde a chuva vem, não se enganem, é verão!!!! (...)”	“Dark and warm, in the afternoon there will be rain; make no mistake, i, it’s summer !!!! (...)”	Before
2016-01-28 13:47:57	“A chuva de ontem @ Em MOEMA https://t.co/3kggiLckGZ”	“Yesterday’s rain @ In MOEMA city https://t.co/3kggiLckGZ”	After
2016-01-27 17:37:39	“Chuva chuva e mais chuva.... https://t.co/vzC9w01qQ8”	“Rain rain and more rain... https://t.co/vzC9w01qQ8”	During

Source: Research data.

Figure 11 – Increasing rainfall and frequency of rain-related geotagged tweets from 1 14:00 BRST to 3 00:00 BRST January 2016, Sao Paulo, Brazil (10-min temporal scale).

Source: Research data.

Figure 12 – Increasing rainfall and frequency of rain-related geotagged tweets from 25 12:00 BRST to 28 08:00 BRST January 2016, Sao Paulo, Brazil (10-min temporal scale).

Source: Research data.

On the basis of this problem statement, the hypothesis raised here is that “there is a lag-time between rainfall measurements and the frequency of rain-related georeferenced social media messages”. In formal terms, the hypothesis can be expressed as follows:

Hypothesis. Let $Q_t = \{q_1, q_2, \dots, q_i, \dots, q_m\}$ and $P_t = \{p_1, p_2, \dots, p_j, \dots, p_n\}$ be defined by two time series, the frequency of rain-related georeferenced social media messages and rainfall data, respectively. The elements q_i and p_j are indexed in time T . For all $t \in T$ there is a constant $k \in \mathbb{Z}$ that makes the relationship function ρ at time t other than zero.

$$\forall_{t \in T}, \exists_{k \in \mathbb{Z}} : \rho(P_t, Q_{t+k}) \neq 0 \quad (3.1)$$

where P_t and Q_{t+k} are observations (values) of the variables at time t and $t + k$, and k is the lagged k -periods. Here, the function ρ is a measure of correlation between both time series. I assessed ρ through the cross-correlation method since it focuses on the delay or separation of time series in terms of peaks and valleys instead of point-to-point alignments assumed by other time series methods, such as dynamic time warping (KRUSKAL; LIBERMAN, 1999). Although the cross-correlation is a simple technique and others more sophisticated within the disciplines of times series and signal processing could be used, it described very well the time differences observed between the content of the Twitter messages and the rain phenomenon through the rain gauge measurements. Therefore, cross-correlation satisfied the conceptual model for mapping the temporal validity between both times series.

Furthermore, an attempt is made to answer the two following questions: “What happens when the k constant is other than zero?” “How many lagged k -periods are needed to provide a suitable description of the social media messages about rain gauge data?” These questions are necessary to show how the temporal approach can be applied, together with the existing approaches, to detect patterns and support the extraction of key information about rainfall events.

3.2 Case study

Context of the study – the rain phenomenon in the city of Sao Paulo, Brazil

The case study was carried out in the city of Sao Paulo, Brazil, where heavy rain often affects the city’s infrastructure and citizens. Figure 13 shows the scenario with the spatial distribution of the rain-related geotagged tweets vis-à-vis the rain gauges and topographic data. As can be seen, the rain-related geotagged tweets are so spatially distributed as the rain gauges.

Figure 13 – Map of the distribution of the rain-related geotagged tweets vis-à-vis the rain gauges and topographic data, in Sao Paulo, Brazil.

Source: Elaborated by the author.

Description of the data and time series

Twitter data

The social media dataset contains 243,333 public geotagged tweets retrieved by Twitter Streaming API from 1 to 30 January 2016. All the tweets selected were geotagged within the administrative borders of the city of São Paulo.

At first, the geotagged tweets collected via Twitter Streaming API were filtered with the aid of keywords and synonyms (including stem words and wildcard). The keywords in Brazilian-Portuguese were “chuv*” (chuva, chuveiro, chuvarada, etc.), “garoa*” (garoando, etc.), “temp*” (temporal, tempestade, tempo ruim, etc.), “alag*” (alagamento, alagado, etc.), and “inund*” (inundação, inundado, etc.). These keywords were chosen and extended from a list of previous studies (ASSIS *et al.*, 2015). After this, a manual search was carried out for the subset of potential rain-related geotagged tweets to remove those false-positive (2,916 geotagged tweets were removed). In this context, the false-positive are those geotagged tweets that contains one or more keywords, but that are not inside the study context; i.e., rain-unrelated geotagged tweets.

For example, there are several rain-unrelated geotagged tweets containing the keyword “garoa” because the city of Sao Paulo is known as “the land of drizzle”.

Furthermore, all retweets messages were removed (228 retweets), i.e., messages concerning of another Twitter user and that contains the text “RT”. With regard to messages with Twitter Emoji (e.g., 🌧️) and grammar mistakes (e.g. “Xuva” instead of “CHuva”), they were classified along with unrelated messages since they did not contain the keywords (239,569 tweets were classified as unrelated). Table 4 shows some examples of rain-related geotagged tweets. In this case study, only 0.25% of the geotagged tweets (620 geotagged tweets) were marked as rain-related tweets. Similar to Chapter 1, such low percentage of geotagged tweets related to a given phenomenon was also observed in other studies of social media analytics (XIAO; HUANG; WU, 2015; HUANG; XIAO, 2015; ALBUQUERQUE *et al.*, 2015).

Table 4 – Examples of rain-related geotagged tweets that have been classified manually.

Date/Time	rain-related tweets	Translation
2016-01-09 01:19:57	“CHUVAAAAA”	“heavy rain”
2016-01-09 15:20:40	“Que chuvinha mais gostosa”	“what wonderful rain”
2016-01-11 07:43:41	“ta garoando aqui”	“it’s drizzling here”
2016-01-27 15:58:40	“Preparando o barco, chove muito em sampa.”	“Get the boats ready, it’s pouring with rain in São Paulo”
2016-01-27 18:56:30	“Tempestade! @ Avenida Ipanema Sorocaba https://t.co/HkvrBIBgeW”	“A Storm! @ Ipanema Sorocaba Avenue https://t.co/HkvrBIBgeW”

Source: Research data.

Rainfall data from the rain gauges

The rainfall data were obtained from the Brazilian National Center of Monitoring and Early Warning of Natural Disaster (CEMADEN) (www.cemaden.gov.br) through a REST API. There were 81 active rain gauges in Sao Paulo in the data collection period. The rainfall measurements were provided in linear depth (millimeters) and with two sizes (intervals) of temporal window: 10 min when it was raining and 60 min otherwise. For this study, only reliable rain gauges were used; those that had missing data throughout the month or day and those that registered inconsistent values were removed. In total, 21 rain gauges were removed.

3.3 Methodology

Figure 14 shows the methodology for analyzing the temporal relationship between rainfall data obtained from rain gauges and Twitter data related to the rain phenomenon. This methodology comprises three steps, which are as follows:

Step 1 To gather sets of rain-related geotagged tweets and rainfall data from rain gauges;

Step 2 To generate a time series of the rainfall data and rain-related geotagged tweets; and

Step 3 To analyze the cross-correlation between both times series of rainfall and the frequency of geotagged tweets related to the rain phenomenon.

Figure 14 – Methodological approach for designing the temporal approach.

Source: Elaborated by the author.

In Step 1, I selected a set of geotagged tweets that were published within the city of Sao Paulo and that contained the previous keywords (i.e., those keywords related to the rain phenomenon). These filtered tweets are called rain-related tweets (Figure 14). Rainfall data were obtained from rain gauges located within the city of Sao Paulo. After this, I generated two time series, i.e., one for the rainfall data and another for the rain-related tweets (Step 2). Both time series were generated with temporal aggregation scales of 10, 20 and 30 min. The rain-related tweets time series corresponds the frequency of rain-related geotagged tweets during the period of analysis, while the rainfall time series are rain gauges data interpolated at the centroid of the city of Sao Paulo using the Inverse Distance Weighting (IDW) method. Finally, I calculated the correlation between them (Step 3) for each temporal scale (10, 20 and 30 min.). The correlation

was used because it can be applied to measure the relationship between two intensities, such as two distinct time series that measure the same natural phenomenon. Step 1 involves manual tasks, whereas Steps 2 and 3 are both automatic.

Methods

The next sections describe the methods related to the spatial interpolation of the rainfall time series (Step 2), the cross-correlation between the time series of rainfall and the frequency of rain-related tweets (Step 3) and how the correlation should be interpreted for this study case.

Inverse distance weighting

Inverse Distance Weighting (IDW) is a technique widely used in hydrological studies for estimating the values of rainfall measurements at the centroid of a catchment by calculating the weighted averages of the available rain gauges (BARTIER; KELLER, 1996; AHRENS, 2006; YANG *et al.*, 2015; MAIR; FARES, 2011). The rain gauges nearest to the centroid will be associated with a greater weight for calculating the average rainfall. The rainfall measurements at the centroid can be calculated as follows:

$$p_c^t = \frac{\sum_{i=1}^N (\frac{1}{d_i^r} p_i^t)}{\sum_{i=1}^N (\frac{1}{d_i^r})} \text{ if } d_i \neq 0, \text{ for all } i \quad (3.2)$$

where p_c^t is a rainfall measurement at the centroid, p_i^t is a rainfall measurement in each rain gauge, d_i is the distance from a rain gauge i to the centroid, and r is the power parameter. In this case study, I considered, for the sake of simplification, that the urban area of the city of Sao Paulo forms a single catchment. Such an assumption can lead to limitations in assuming uniform global patterns of rain that normally do not occur over large areas as the city of Sao Paulo. Figure 15 illustrates the spatial interpolation calculated for the rainfall data (using the parameter $r = 2$).

Cross-correlation

Let q_t and p_t be two stationary processes. The correlation is any statistical relationship, whether causal or otherwise, between two random variables or two sets of data. Cross-correlation is already a function that estimates the correlation between q_t and p_t at pairs of time points (BACCHI; KOTTEGODA, 1995). The cross-correlation at lag k is given as follows:

$$\rho_{qp}(k) = \frac{\sum_{t=1}^N (q_{t+k} - \mu_q)(p_t - \mu_p)}{\sqrt{\sum_{t=1}^N (q_t - \mu_q)^2 \sum_{t=1}^N (p_t - \mu_p)^2}} \quad (3.3)$$

where μ_q and μ_p are the expected values of q and p , respectively, and k is a lag of the variable q . Cross-correlation is a method to check if the two series have randomness. Such a randomness

Figure 15 – Spatial interpolation of the active rain gauges in the city of Sao Paulo, Brazil.

Source: Elaborated by the author.

is ascertained by computing autocorrelations for data values at varying lag-times. If they are random, the correlations should be near zero for any and all lag separations. Otherwise, one or more of the correlations will be significantly non-zero. With k negative, they are predictors of p_t , it is sometimes said that q leads to p . When one or more q_{t+k} , with k positive, are predictors of p_t , it is sometimes said that q lags p .

3.4 Results

Temporal scale of 10-min

As can be seen from [Figure 11](#) and [Figure 12](#) (see [Section 3.1](#)), the time series of rainfall and geotagged tweets related to the rain phenomenon are not synchronized, i.e., there is not an exact correspondence between both the time series. This evidence appears throughout the case study. Another piece of evidence is that the peaks of both time series sometimes shift in time, i.e., the peak of the frequency of geotagged tweets related to the rain phenomenon appears before the peak of the rainfall measurements and vice-versa. For example, from 20:00 BRST to 21:00 BRST

on 1 January 2016 the frequency of geotagged tweets related to the rain phenomenon came before the rainfall measurements (Figure 11), but from 17:00 BRST to 18:00 BRST on 2 January 2016 the frequency of rain-related geotagged tweets came after the rainfall measurements (Figure 12). A similar kind of behavior can be observed in other periods in January 2016.

Figure 16 depicts the cross-correlation for the period from 1 14:00 BRST to 3 00:00 BRST January 2016, whereas Table 5 summarizes the cross-correlation values for the same period. The blue dotted line represents the confidence interval of 95%.

Figure 16 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 14:00 BRST to 3 00:00 BRST January 2016.

Table 5 – Cross-correlation values between rainfall data time series and rain-related tweets with a temporal scale of 10-min for the period from 1 14:00 BRST to 3 00:00 BRST January 2016.

Lag-unit	Cross-correlation	Lag-unit	Cross-correlation
-10	0.065	1	0.451
-9	0.078	2	0.328
-8	0.135	3	0.264
-7	0.174	4	0.245
-6	0.246	5	0.210
-5	0.220	6	0.218
-4	0.221	7	0.207
-3	0.247	8	0.091
-2	0.378	9	0.056
-1	0.563	10	0.116
0	0.528	-	-

Source: Research data.

Although the significance is within the range of -7 to 7 lag-units, the greatest correlation value is -1 lag-unit ($k = -1$). This means most rain-related geotagged tweets precede the rain phenomenon in 10 min since each lag-unit refers to a temporal interval of 10 min. The most likely explanation is that rain-related geotagged tweets can ‘predict future rain episodes’ or detect the rain phenomenon in near real-time, i.e., people give a forecast of the rain phenomenon (see Table 3). Therefore, such a time window setting can be used for forecasting and monitoring the rain phenomenon and, thus, provide better support for decision-making.

A similar result is achieved for the period from 25 12:00 BRST to 28 08:00 BRST January 2016 (Figure 17), i.e., the greatest correlation value is -1 lag-unit ($k = -1$). In this case, the confidence interval (Figure 17) and cross-correlation values (Table 6) are highest than the previous period (Figure 16).

Figure 17 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 25 12:00 BRST to 28 08:00 BRST January 2016.

In fact, the results show that the highest correlation is negative ($k = -1$), but the positive correlation value of 1 lag-unit ($k = 1$) could be of great value for monitoring activities. For example, the decision-maker might monitor what the social media users are saying during and right after an episode of heavy rain and assess the consequences and effects of that heavy rain episode (e.g., situational updates about an ongoing event). Thus, this setting ($k = 1$) is useful for monitoring extreme events, such as heavy rainfall episodes. Hence, the lag-time can be used to define a threshold time for monitoring the affected area. For example, a maximum size of a temporal window of 20 min can be defined for getting information after the event has been detected, i.e., $0 < k < 2$.

Table 6 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 25 12:00 BRST to 28 08:00 BRST January 2016.

Lag-unit	Cross-correlation	Lag-unit	Cross-correlation
-10	0.140	1	0.633
-9	0.151	2	0.505
-8	0.178	3	0.365
-7	0.207	4	0.273
-6	0.264	5	0.184
-5	0.340	6	0.186
-4	0.457	7	0.259
-3	0.594	8	0.345
-2	0.712	9	0.394
-1	0.772	10	0.350
0	0.744	-	-

Source: Research data.

Another possible setting is when the lag-unit is zero ($k = 0$). In this case, rainfall data and rain-related geotagged tweets could be analyzed for monitoring activities, either together or separately. While the first alternative is suitable to improve the analysis in (near) real-time, the second alternative could be applied when the rainfall data are insufficient. In this case, it is necessary to assess the credibility of the social media messages. However, credibility issues in social media analysis are outside the scope of this doctoral research.

Figure 18 shows the results for the entire period (from 1 to 30 January 2016). It can be seen that the cross-correlation is greater than 0.5 for the lag-units that varies from -1 to 0 (Table 7). In brief, the -1 lag-unit setting (-10 min) is enough to be used in forecasting rainfall, whereas the 0 lag-unit (0 min) can be used in a (near) real-time setting.

Table 7 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 to 30 January 2016.

Lag-unit	Cross-correlation	Lag-unit	Cross-correlation
-10	0.141	1	0.445
-9	0.143	2	0.383
-8	0.165	3	0.302
-7	0.185	4	0.241
-6	0.220	5	0.179
-5	0.267	6	0.188
-4	0.341	7	0.214
-3	0.413	8	0.247
-2	0.499	9	0.270
-1	0.548	10	0.253
0	0.518	-	-

Source: Research data.

Figure 18 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 10-min for the period from 1 to 30 January 2016.

Source: Research data.

Temporal scales of 20 and 30-min

Similar results were obtained for the entire period of the case study with temporal scales of 20 and 30-min (Figure 19 and Figure 20). Nevertheless, these results reveal additional findings related to the temporal aggregation effect of MTUP (CHENG; ADEPEJU, 2014).

As can be seen from Table 7, Table 8 and Table 9, the cross-correlation value for the 0 lag-unit increases as the temporal resolution increases. In this way, the results indicate that a large temporal scale affect the cross-correlation, by making both time series converge to an “exact correspondence”. In other words, the temporal aggregation process may lead to a false view of the scenario for decision-making. For example, Twitter messages that precede rain events in terms of their metadata (e.g., content and timestamp) may fall mistakenly within the time interval of the event (i.e., from before to during the rain phenomenon).

Figure 21 shows the temporal effect of aggregation by comparing the cross-correlation values for the temporal scales of 10, 20, and 30-min and period from 1 to 30 January 2016. The cross-correlation is affected by the temporal aggregation of the time series since the number of resulting intervals (i.e., time bins) is reduced (CHENG; ADEPEJU, 2014). Such a temporal aggregation process shows that large temporal resolutions are not suitable for detecting and monitoring rainfall-events in (near) real-time.

Figure 19 – Cross-correlation and visualization between the times series of rainfall data and rain-related geotagged tweets with a temporal scale of 20-min for the period from 1 to 30 January 2016.

Source: Research data.

Table 8 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 20-min for the period from 1 to 30 January 2016.

Lag-unit	Cross-correlation	Lag-unit	Cross-correlation
-10	0.108	1	0.454
-9	0.133	2	0.267
-8	0.130	3	0.215
-7	0.165	4	0.297
-6	0.158	5	0.302
-5	0.171	6	0.259
-4	0.193	7	0.279
-3	0.260	8	0.267
-2	0.406	9	0.249
-1	0.586	10	0.185
0	0.601	-	-

Source: Research data.

Figure 20 – Cross-correlation and visualization between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 30-min for the period from 1 to 30 January 2016.

Source: Research data.

Table 9 – Cross-correlation values between the time series of rainfall data and rain-related geotagged tweets with a temporal scale of 30-min for the period from 1 to 30 January 2016.

Lag-unit	Cross-correlation	Lag-unit	Cross-correlation
-10	0.085	1	0.425
-9	0.110	2	0.309
-8	0.114	3	0.313
-7	0.123	4	0.301
-6	0.143	5	0.324
-5	0.169	6	0.294
-4	0.184	7	0.195
-3	0.209	8	0.122
-2	0.319	9	0.093
-1	0.581	10	0.051
0	0.637	-	-

Source: Research data.

Figure 21 – Comparison of cross-correlation between the time series of rainfall data and rain-related geotagged tweets with temporal scales of 10, 20 and 30-min for the period from 1 to 30 January 2016.

Source: Research data.

3.5 Discussion

The results provide evidence that the geotagged tweets related to the phenomenon and rainfall measurements are associated with different lag-times (Figures 16, 17 and 18), but they are more associated with lag-times that ranges from -10 to +10 minutes (Table 5, Table 6 and Table 7). Statistically, the lag-time can vary from negative to positive (see Table 5, Table 6 and Table 7). This temporal characteristic can be hard to detect with the human eye without a temporal correlation. For example, when the analyst chooses a coarse temporal scale, the peaks of rainfall and frequency of rain-related geotagged tweets tend to be very close over time; otherwise, they tend to present large or small differences as a function of temporal aggregation. Thus, it is not clear to an analyst whether the rain-related geotagged tweets are or not useful for decision-making. However, the results show that rainfall patterns can be approximated by rain signals extracted from Twitter data when take the duration, frequency, and lag-time into account. Such temporal components, thus, should guide the development of monitoring tools that make use of data from the social media as a supplementary data source.

Moreover, the temporal approach can be applied to detect rainfall patterns in real-time using official data and Twitter data. In addition, the findings suggest that this approach can be useful to fit or approximate the best temporal scale between two time series in any setting (e.g., real-time, near real-time, and post-hoc analysis) since individual data sources can be represented from single time series. This is useful for monitoring activities, where the scale and frequency of the data can change in space and time.

Thus, the main value of this contribution is that I have put forward a novel approach to describe the temporal relationship between rainfall data and rain-related social media data using lag-times rather than only relying on the duration and frequency of the time series. This opens up a new way of exploring and improving the temporal models and approaches with the aim of creating functional relations to enhance hydrological modelling for monitoring rainfall in real-time. Moreover, it is possible to improve existing detection systems, such as PrioritizeSN ([ASSIS *et al.*, 2015](#)), Toretter ([SAKAKI; MATSUO, 2012](#)), GeoCONAVI ([SPINSANTI; OSTERMANN, 2013](#)), to name just a few.

One limitation of this approach that describes the time differences between the time series of rainfall and frequency of rain-related geotagged tweets, however, is the shortcoming of the cross-correlation technique in capturing seasonal and causal effects. In the future, such effects should be investigated to complement the temporal analysis by showing more evidence about the use of social media data as a supplementary data source for rainfall monitoring, given the correlation has shown initial evidence about it.

3.6 Final remarks

People report rain-related messages on social media platforms in synchronous and asynchronous ways in relation to a rain event that they have observed or heard about. Since such messages can be reported before, during, and after the occurrence of a rain event, some of them are either outdated or uncertain for decision-making – or to generate rain signals from social media data. In this chapter, I argued that the relationship between rain signals extracted from social media data and the rain phenomenon should be analyzed by means of lag-times rather than only the duration and frequency. Thus, a temporal approach relied on the cross-correlation method was established for the identification of valid lag-times in this relationship.

Finally, future work should further extend this approach by incorporating other social media platforms (e.g., Instagram and Flickr) and case study scenarios (e.g., other cities and countries) to be able to obtain a generalization.

A CONCEPTUAL SPECIFICATION MODEL

Social media activity is spatially heterogeneous in nature (ZHOU; CHEN, 2014). People interact through social media platforms at different periods and places in cities, leading to clusters of high densities of messages in some places and areas that have low densities of messages in between. These spatially discontinuous places that share high densities of messages reveal the spatial heterogeneity of social media activity, which has been explained by spatio-temporal externalities related to urban environment. Thus, the variability of social media activity is not distributed randomly over the city.

Most studies have explored the spatial heterogeneity of social media activity through the geo-demographic context. In short, they sought to correlate the overall social media activity with demographic and economic factors (SLOAN, 2017; LONGLEY; ADNAN, 2016; XIAO; HUANG; WU, 2015; LI; GOODCHILD; XU, 2013). Nonetheless, these factors describe the population dynamics from the perspective of where people live (i.e., people's homes) rather than how they move around the city. At the same time, social media activity is a set of snapshots of people's everyday behaviours, including their footprints in cities (CHENG *et al.*, 2011). Thus, scant attention is given to explain both social media activity and its spatial heterogeneity from different perspectives than those related to demographic and socioeconomic factors, such as urban mobility.

Indeed, human mobility data is a rich data source of migratory flows at different levels of spatial analysis (i.e., micro-, meso-, and macro-level) and has recently been studied through social media data (RODRÍGUEZ *et al.*, 2020; JURDAK *et al.*, 2015). Even faced by sampling biases, such as the amount of georeferenced data and low representation of the real population, social media data is capable of inferring inter- and intra-zonal flows in cities (RODRÍGUEZ *et al.*, 2020). Thus, human mobility data and social media data are spatially correlated and, hence, the former could contribute to explain the spatial patterns of the latter – especially for very heterogeneous social media activity, as those related to a particular topic (e.g., heavy rain).

This chapter establishes a linear regression model combining demographic, socioeconomic, and human mobility data to characterize the places where social media activity accurately mirrors rain patterns. Linear regression has been a useful and consistent tool to explore the relationship between a dependent variable and one or more explanatory variables, including in urban spatial analyses (HAINING, 2003; SINGLETON; SPIELMAN; FOLCH, 2018).

A linear regression model is applied in a case study in Sao Paulo, Brazil, and advances in determining the likely values of the exploratory variables to enable regions to accurately mirror the rain patterns. The remainder of this chapter is structured as follows: Section 4.1 provides an overview of the literature on the spatial heterogeneity in social media data; Section 4.2 presents the materials and methods, including the study area; Section 4.3 outlines the conceptual model and an exploratory analysis of the exploratory variables; Section 4.4 reports the main results, while Section 4.5 conducts a discussion on the findings; finally, Section 4.6 presents the conclusions and makes suggestions for future studies.

4.1 The spatial heterogeneity in social media data

Spatial heterogeneity is related to “the lack of stability over space of the behavioral or other relationships under study . . . and this implies that functional forms and parameters vary with location and are not homogeneous throughout the dataset” (ANSELIN, 1988, pp. 9). This variation over geographic space may be even more complex in social media data than in other datasets, since spatial heterogeneity here is caused by bias in the production practices of social media users and data acquisition process and varies in different types of social media platforms (RZESZEWSKI, 2018; WESTERHOLT *et al.*, 2016). This means social media data have an uneven spatial distribution over geographic space.

Figure 22 shows this uneven spatial distribution in Twitter data acquisition over the districts of the city of Sao Paulo, Brazil. As can be seen from Figure 22 (a), it is reasonable to consider a power law (fit with an exponent value of 2.0) as a possible model for describing the density of tweets over the districts. We can note that only 1% of districts present more than 100,000 tweets, while 50% of the districts present less than 10,000 tweets (Figure 22 (b)). This probability distribution has also been observed in other studies in social media analytics (JIANG; MIAO, 2015; WANG *et al.*, 2018; RZESZEWSKI, 2018).

Beyond the power law distribution, another characteristic shown in Figure 22 (a) concerns the (dis)contiguity of the districts with similar densities of tweets. This characteristic stands for the spatial dependence of Twitter density over space and is very often quantified by using spatial autocorrelation measures like Global Moran’s I coefficient. In Figure 22 (a), Global Moran’s I read off a value of 0.18, which means similar Twitter densities are weakly clustered over space. Even when we change the spatial arrangement of the study area (i.e., shape and size of areal units, extent of the study area, etc.), similar observations are obtained.

Figure 22 – (a) Aggregation of geotagged tweets into districts of the city of Sao Paulo and (b) the power law distribution of geotagged tweets. Both for the period from 7 to 30 November 2017.

Source: Elaborated by the author.

While both power law rule and global spatial dependence are thus complementary helpful tools to understand the spatial distribution and identify the spatial heterogeneity of social media activity, they do not serve to explain what factors yields the uneven distribution of social media data production. This has limited the capacity of use social media data in practical solutions such as crisis and emergency situations ([IMRAN *et al.*, 2015](#)) and urban planning ([NUMMI, 2017](#)). Because of this, a lot of research have investigated what factors influence the production of social media data.

Most of the studies have reported that social media data as strongly influenced by spatial externalities related to demographic and socioeconomic factors, and cultural and ethnic groups. For example, [Li, Goodchild and Xu \(2013\)](#) found a linear relationship between tweet density and the percentage of well-educated people with an advanced degree and a good salary, as well as between photo density from Flickr and the percentage of white and Asian people in the United States. [Sloan \(2017\)](#) found that Twitter users are predominantly young in comparison to the UK population as of the 2011 Census and that certain occupational groups are more likely to use Twitter than others. [Longley and Adnan \(2016\)](#) found that areas concentrated in Central London are connected and yield higher volume of tweets than most outer areas. In summary, younger, more affluent, and well-educated people in places with better infrastructure are more likely to be involved in the generation of social media data.

In extension to linear relationships found in the previous works, [Xiao, Huang and Wu \(2015\)](#) found a reversed U-shaped relationship between the number of Hurricane Sandy-related tweets and the measures of income and median housing value; i.e., the participation of social media users increases with increases in wealth, but until a turning point when the participation decreases. They argued this nonlinear relationship reflects both the 'digital divide' (i.e., the gap

between people who do and do not have access to information and communication technologies) and motivation (or lack thereof) among social classes (wealthy/poor). Despite that, they also found a linear relationship with the residential population size; except in landmark sites (e.g., airports and green spaces) where the number of messages is expected to be inflated.

Although previous studies have advanced at understanding the spatial heterogeneity of social media activity, demographic and socioeconomic factors only describe the population dynamics from the perspective of where people live (i.e., people's homes). At the same time, social media activity is a set of snapshots of people's everyday behaviours and, thus, should be better captured from the perspective of how people move around the city. Thus, the investigation of the uneven spatial distribution of social media activity should take other perspectives into account, such as human mobility. In addition, most of the studies investigate the spatial distribution of the overall social media activity, while 'thematic social media data' has been marginally addressed (i.e., those data around a topic that is capable of revealing the human perception against a phenomenon of interest). Social media data around a topic can introduce higher levels of heterogeneity than the overall social media data, since it is more scattered within the urban space. In light of this, this work carries out a case study to understand the spatial distribution and characterize the places where social media activity accurately mirrors rain patterns.

4.2 Materials and methods

Study area

The study area was the city of Sao Paulo, Brazil, which has received much attention due to heavy rain events that cause flash floods, inundations and flooding, mainly due to the rivers overflowing and the poor drainage system of the urban pavements. It should also be noted that Sao Paulo is an important economic centre and has a vast number of Twitter users and an estimated population of approximately 12 million people of different ethnics, nationalities and social classes (IBGE, 2010), which has made it the most density populated city in Brazil and an interesting place to geodemographic studies. Another reason that makes Sao Paulo the study area is the publicly available data about inter- and intra-zonal flows and rainfall rates.

Datasets

2010 Census data. Demographic and socioeconomic data were obtained from the 2010 Brazilian Census Survey, which is provided and maintained by the Brazilian Institute of Geography and Statistics (IBGE). This survey is complete and tracks 320 census tracts within the city of Sao Paulo and more than 200 variables; but, only population, income, and level of education were selected as explanatory variables for this study. The selection of these variables was assisted by Section 4.1 and the assumptions of the conceptual model (Section 4.3).

Human mobility data. The mobility data came from the 2017 Metro Subway Origin & Destination Survey for the Sao Paulo Metropolitan Region (SPMR), which has been conducted every 10 years since 1967 (SPMR, 2017). This survey divides the city into 460 traffic zones and includes more than forty-two million commuting flows, as well as variables relating to the number of trips between zones, purpose (e.g., work and school), origin and destination, and travel mode (i.e., motorized or non-motorized, public or private, and walk, bicycle, subway, etc.). For this study, I considered the number of commuting flows (i.e., the number of inflows and outflow of the zones) independently of the purpose and mode of transport. Furthermore, a vector aggregating the number of inflows in each spatial unit was summarized from the O&D matrix.

Twitter data. Twitter data used in this study includes a sample of 2,073,219 million public geotagged tweets collected within the city of Sao Paulo during the period from November 2016 to November 2017 from Twitter Streaming API. This amount of geotagged tweets corresponds to approximately 7.5% of the total number of tweets collected for the entire period. Although Twitter Streaming API returns a large enough set of geotagged tweets (MORSTATTER *et al.*, 2013), such a sample may not be representative of the census population (SLOAN, 2017). Such a potential lack of representativeness was assumed as a bias in the case study.

Rainfall data. Rainfall data were obtained from the weather radar measurements of the DAEE and the FCTH of the Polytechnic School of the University of Sao Paulo. This device is a Dual Polarisation Doppler S-Band weather radar located approximately 60-70 km far from Sao Paulo city and continuously estimates the statistics of rainfall rates at a 250 meter spatial resolution from an azimuthal width of 1 degree (SELEX, 2015). Instead of using the polar rain yields, a radar product known as SRI projected in Cartesian coordinates of a 500x500 meter scale was relied on (SELEX, 2015). This enabled us to create new maps to match the representation of rainfall rates used in the Twitter activity. The mean value of each time interval was stored in each areal unit of analysis. Spatial and linear interpolation techniques were employed to overcome the problem of missing data for measurements within the same day.

Methods

Filtering geotagged tweets related to the heavy rain phenomenon. The geotagged tweets were filtered based on five Brazilian keywords related to the rain phenomenon ('chuva', 'chove', 'chuvoso', 'chuvosa' and 'chuvarada'), in accordance with the filtering process described in Section 2.3 (Chapter 2). Such keywords were obtained from Andrade *et al.* (2018a) and filter more true-positive than true-negative rain-related geotagged tweets. True-negative tweets are referred to as those that contain at least one keyword and where the text content is not linked to the rain phenomenon. An example of a true-negative tweet is 'chuva de arroz após a cerimônia de casamento' (rice rain after wedding ceremony). While this filtering strategy prioritizes the precision over recall, it does not consider false-positive rain-related geotagged tweets, such as 'toró d'água' (storm).

Selecting the areal unit of analysis. Since Twitter activity is not driven by census tracts or any other administrative boundary, the entire surface area of the city was partitioned into hexagonal areal units of 30 km². This areal unit was selected based on the multicriteria optimization framework (ANDRADE *et al.*, 2020), which minimizes the total amount of variance of the rain-related geotagged tweets patterns. Each spatial areal unit aggregated rainfall data and rain-related geotagged tweets over the entire period of analysis – from November 2016 to November 2017 – on a one-day temporal scale. Following the procedures in Chapter 2, I considered only those days classified as heavy shower (i.e., days with the daily record of rainfall equal to or greater than 10 mm per hour. In addition, the ‘odds ratio measure’ of the frequency of geotagged tweets was used for this case study – further details how this measure was calculated see Equation 2.2 (Chapter 2).

Calculating the intensity of the social mirroring of rain patterns from Twitter data. This intensity corresponds to the cross-correlation measure (at the lag-unit 0) between the time series of odds ratio measure of the frequency of geotagged tweets and rainfall rates. In general, cross-correlation is a measure of similarity of two time series as a function of the displacement of one relative to the other – further details how this measure was calculated for each spatial unit see Equation 3.3 in Chapter 3. Although the displacement between both times series varies from -1 to +1 lag-units (see Section 3.4), I selected the lag-unit where the cross-correlation is highest for the largest number of spatial units of analysis (lag-unit = 0). It is important to note this work does aim at studying where social media activity accurately mirrors rain patterns and the influence of demographic, socioeconomic, and human mobility factors. Hence, the premise here is that spatial units exhibiting high correlation are more likely to use the aggregation of geotagged tweets related to the rain phenomenon as a proxy variable of rainfall rates.

4.3 The conceptual specification model

The intensity of the social mirroring of rain patterns extracted from Twitter data was investigated through an ordinary least squares (OLS) regression model, since no significant spatial effects were diagnosticated applying the Lagrange Multiplier test (LM test) as suggested by Anselin (1988).

In line with the narrative in Section 4.1, a high intensity of social mirroring of rain patterns is expected in the spatial units characterized by high levels of population, income, and well-educated people. Spatial units exhibiting these characteristics are, thus, more likely to accurately mirror the rain patterns from Twitter activity. Alongside the variables population, income, and well-educated, the number of people that arrive in the spatial units is also expected to explain such intensities. This assumption is based on previous studies that reported Twitter data as a reliable, but limited, source to study human mobility patterns (RODRÍGUEZ *et al.*, 2020; JURDAK *et al.*, 2015) and whereabouts of people (STEIGER *et al.*, 2015). Thus, how people commute among places within the city should partially explain the social media activity.

In summary, it is expected that the population, income, and the number of inflows from the O&D matrix should be positively correlated with the social mirroring of rain patterns extracted from Twitter data. It should be noted that the well-educated variable was removed from the Equation 4.1 since it shares similar spatial patterns with the population (Figure 24).

The OLS regression model specification is as follows.

$$y = \beta_0 + \beta_1 \text{population} + \beta_2 \text{income}_{(avg)} + \beta_3 \text{inflow} \quad (4.1)$$

where y is the intensity of the social mirroring of rain patterns from Twitter data over the spatial units, population is the number of people that live in the spatial units, and inflow corresponds to the number of trips that arrives into the spatial units.

The number of trips arriving into the spatial units was derived from the O&D matrix by reshaping the traffic zones into the spatial units of 30 km² and, later, aggregating the rows (origins) by columns (destinations) – the aggregation vector determines the number of inflow in the spatial units of the reshaped zoning system. Since both zonal systems are incompatible and the data transforming process faces the polygonal overlay problem (GOTWAY; YOUNG, 2002) (i.e., the spatial units of 30 km² might totally or partially cover the traffic zones), the percentage of overlapping areas was used as weights to estimate the inflow/outflow of the reshaped zoning system or reshaped O&D matrix. Figure 23 shows this reshaping process between two arbitrary and incompatible zoning systems.

Figure 23 – Example of reshaping between two arbitrary and incompatible zoning systems. (a) two incompatible zoning systems (1–4 and A,B). (b) the original inflow/outflow matrix (or Origin-Destination matrix). (c) the weight matrix based on the overlap between the areas of the zoning systems. (d) the overlapping equation. (e) the calculation of the aggregation vector containing the number of inflows for each area of the reshaped zoning system.

(a)

$$X = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \end{matrix}$$

(b)

$$W = \begin{matrix} & \begin{matrix} A & B \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0.25 & 0 \\ 0 & 0 \\ 0.25 & 0.75 \\ 0 & 0.38 \end{bmatrix} \end{matrix}$$

(c)

$$x_{kl} = \sum_{i=1}^m \sum_{j=1}^m x_{ij} w_{ik} w_{jl}, \quad k, l = \{A, B\} \quad i, j = \{1, 2, 3, 4\} \quad m = 4$$

(d)

Note – Continued on next page.

$$\begin{array}{c}
 \begin{array}{cc}
 & A \\
 A & \left[\begin{array}{l}
 (x_{11}w_{1A}w_{1A}) + (x_{12}w_{1A}w_{2A}) + (x_{13}w_{1A}w_{3A}) \\
 + (x_{14}w_{1A}w_{4A}) + (x_{21}w_{2A}w_{1A}) + (x_{22}w_{2A}w_{2A}) \\
 + (x_{23}w_{2A}w_{3A}) + (x_{24}w_{2A}w_{4A}) + (x_{31}w_{3A}w_{1A}) \\
 + (x_{32}w_{3A}w_{2A}) + (x_{33}w_{3A}w_{3A}) + (x_{34}w_{3A}w_{4A}) \\
 + (x_{41}w_{4A}w_{1A}) + (x_{42}w_{4A}w_{2A}) + (x_{43}w_{4A}w_{3A}) \\
 + (x_{44}w_{4A}w_{4A})
 \end{array} \right. \\
 B & \left[\begin{array}{l}
 (x_{11}w_{1B}w_{1A}) + (x_{12}w_{1B}w_{2A}) + (x_{13}w_{1B}w_{3A}) \\
 + (x_{14}w_{1B}w_{4A}) + (x_{21}w_{2B}w_{1A}) + (x_{22}w_{2B}w_{2A}) \\
 + (x_{23}w_{2B}w_{3A}) + (x_{24}w_{2B}w_{4A}) + (x_{31}w_{3B}w_{1A}) \\
 + (x_{32}w_{3B}w_{2A}) + (x_{33}w_{3B}w_{3A}) + (x_{34}w_{3B}w_{4A}) \\
 + (x_{41}w_{4B}w_{1A}) + (x_{42}w_{4B}w_{2A}) + (x_{43}w_{4B}w_{3A}) \\
 + (x_{44}w_{4B}w_{4A})
 \end{array} \right.
 \end{array}
 \end{array}
 \begin{array}{c}
 \begin{array}{cc}
 & B \\
 A & \left[\begin{array}{l}
 (x_{11}w_{1A}w_{1B}) + (x_{12}w_{1A}w_{2B}) + (x_{13}w_{1A}w_{3B}) \\
 + (x_{14}w_{1A}w_{4B}) + (x_{21}w_{2A}w_{1B}) + (x_{22}w_{2A}w_{2B}) \\
 + (x_{23}w_{2A}w_{3B}) + (x_{24}w_{2A}w_{4B}) + (x_{31}w_{3A}w_{1B}) \\
 + (x_{32}w_{3A}w_{2B}) + (x_{33}w_{3A}w_{3B}) + (x_{34}w_{3A}w_{4B}) \\
 + (x_{41}w_{4A}w_{1B}) + (x_{42}w_{4A}w_{2B}) + (x_{43}w_{4A}w_{3B}) \\
 + (x_{44}w_{4A}w_{4B})
 \end{array} \right. \\
 B & \left[\begin{array}{l}
 (x_{11}w_{1B}w_{1B}) + (x_{12}w_{1B}w_{2B}) + (x_{13}w_{1B}w_{3B}) \\
 + (x_{14}w_{1B}w_{4B}) + (x_{21}w_{2B}w_{1B}) + (x_{22}w_{2B}w_{2B}) \\
 + (x_{23}w_{2B}w_{3B}) + (x_{24}w_{2B}w_{4B}) + (x_{31}w_{3B}w_{1B}) \\
 + (x_{32}w_{3B}w_{2B}) + (x_{33}w_{3B}w_{3B}) + (x_{34}w_{3B}w_{4B}) \\
 + (x_{41}w_{4B}w_{1B}) + (x_{42}w_{4B}w_{2B}) + (x_{43}w_{4B}w_{3B}) \\
 + (x_{44}w_{4B}w_{4B})
 \end{array} \right.
 \end{array}
 \end{array}
 \end{array}
 \begin{array}{c}
 \begin{array}{cc}
 & A \\
 A & \left[\begin{array}{cc}
 0.125 & 0.565 \\
 0.377 & 1.276
 \end{array} \right] \\
 B &
 \end{array}
 \begin{array}{cc}
 & B \\
 A & \left[\begin{array}{cc}
 0.125 + 0.377 & 0.565 + 1.276
 \end{array} \right] \\
 B &
 \end{array}
 \begin{array}{cc}
 & B \\
 A & \left[\begin{array}{cc}
 0.502 & 1.841
 \end{array} \right] \\
 B &
 \end{array}
 \end{array}
 \end{array}
 \begin{array}{c}
 (e)
 \end{array}$$

The exploratory data analysis

An initial exploratory data analysis is displayed in Figure 24, in which the spatial patterns of the exploratory variables of the conceptual model are summarized. While the population is concentrated in the outer districts (Figure 24 (a)), the income is concentrated in the districts around the city center (Figure 24 (d)). Nonetheless, the patterns of the relation between the number of tweets and population is higher around the city center than outer districts where population is greater (Figure 24 (b)). Indeed, Sao Paulo has an economic human activity that starts high in the center (i.e., Sé district) and smoothing to low as it moves away to the outer districts. On the other hand, the number of trips that arrives into the districts is more spatially spread compared to the previous variables (Figure 24 (d)). Figure 24 (d) reinforces the potential of the human mobility data to explain the social media activity which is heterogeneous in space and spread in cities.

Figure 24 – Quantile maps of the (a) population, (b) number of tweets by population, (c) income, and (d) well-educated.

Source: Elaborated by the author.

4.4 Results

The intensity of the social mirroring of rain patterns from Twitter data

Figure 25 (a) shows the spatial distribution of the intensity of the social mirroring of rain patterns extracted from Twitter data. There are significant differences throughout the spatial units, concentrically varying from the center, with high values, to the outer areas, with low values. At first glance, such an intensity tends to be strongly skewed towards the income (Figure 25 (c)), however, it is not completely true and the OLS regression analysis shows the impact of the population (Figure 25 (b)) and number of trips that arrives into the spatial units (Figure 25 (d)). The regions sharing no data correspond to either poorly urbanized or environmental protection areas. For example, the Cantareira Park is located in the north, while in the south has a mix of non-urbanized and environmental preservation areas.

Figure 25 – Quantile maps of the (a) intensity of the social mirroring of rain patterns, (b) population, (c) income, and (d) inflow.

(a)

Note – Continued on next page.

Source: Elaborated by the author.

OLS regression analysis

Results from the OLS regression yielded strong and consistent results (Table 10). All estimated coefficients are positive and statistically significant, matching with the conceptual model specification discussed in Section 4.3; i.e., the spatial units sharing large population, income average, and number of inflows mirror accurately rain patterns extracted from Twitter data. The regression achieved a suitable fit in terms of adjusted r-square ($R^2 = 0.72$), while the residual distribution was spatially random (Figure 26) and slightly skewed to the right (skewness measure of 0.37). Global Moran's I coefficient of the residuals using a row-standardized spatial weights matrix based on first-order rook contiguity revealed small and non significant negative spatial correlation of -0.10 (p-value = 0.82) (Figure 26). When individually examining the influence of each variable of the conceptual specification model, population, income, and inflow explain 40%, 51%, and 60% of variability, respectively.

The references #1 and #2 correspond to supplementary models established based on Section 4.1, where a number of researchers argue that population, income, and well-educated variables are more likely to generate social media feeds. While reference model #2 explain less the variability of the intensity of the social mirroring of rain patterns ($R^2 = 0.68$), the reference model #1 explain the variability equally to the model ($R^2 = 0.72$). However, the estimated value of the population's coefficient does not match with the conceptual model specification, i.e., the intensity of the social mirroring of rain patterns and population are inversely proportional.

Table 10 – OLS regression statistic results.

Coefficients	Model	Reference model #1	Reference model #2
population	6.344e-07 ($\pm 1.455\text{e-}07$)***	-2.030e-05 ($\pm 7.047\text{e-}06$)**	7.726e-07 ($\pm 1.479\text{e-}07$)***
income _(avg)	8.368e-05 ($\pm 3.446\text{e-}05$)*	9.823e-05 ($\pm 3.049\text{e-}05$)**	1.604e-04 ($\pm 2.407\text{e-}05$)***
well-educated	-	2.644e-05 ($\pm 8.840\text{e-}06$)**	-
inflow	2.364e-08 ($\pm 8.074\text{e-}09$)**	-	-
Adj. R ²	0.72	0.72	0.68
AIC	-64.07	-64.42	-57.53

Note – Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

Source: Research data.

Figure 26 – Residual map of the conceptual specification model (Equation 4.1).

Source: Elaborated by the author.

Determination of arbitrary thresholds

Since the explanatory variables are positively and significantly correlated with the dependent variable (Table 10), the analyst can determine which spatial units are reliable to use Twitter data for decision making through empirical thresholds and scatterplots. Figure 27 shows a piece of evidence in which to obtain a minimum mirroring rate of 40% (cross-correlation = 0.4), the population, income, and inflow variables should be greater than or equal to 31,786 (Figure 27 (a)), 514 (Figure 27 (b)), and 484,235 (Figure 27 (c)), respectively. Thus, these values can determine, to some extent, places where Twitter data accurately mirror rain patterns.

Figure 27 – Scatterplot between the intensity of social mirroring of rainfall patterns and the independent variables of (a) population, (b) income, and (c) inflow.

Source: Research data.

4.5 Discussion

A large body of studies has linked younger, more wealthy, and well-educated people in places with better infrastructure to the production of social media data (SLOAN, 2017; LONGLEY; ADNAN, 2016; XIAO; HUANG; WU, 2015; LI; GOODCHILD; XU, 2013). While these studies have found a high correlation between these demographic and socioeconomic factors and social media activity, they don't take into account how people move around the city. However, social media activity is associated with the human movement in cities and has recently been used to capture rich features of human mobility, such as migratory urban flows (RODRÍGUEZ *et al.*, 2020; JURDAK *et al.*, 2015). Thus, a limitation of these studies that use only demographic and socioeconomic factors is they rest on the assumption that the highest intensities of social media activity occur where people live rather than the places where they usually go or passing by. In light of this, through human mobility data, it is possible to extend the previous exploratory models to improve the characterization of the spatial patterns of urban social media activity.

In this chapter, I examine human mobility data as an independent variable of the number of trips that arrive in the spatial units to characterize the places where social media activity accurately mirror rain patterns. An OLS regression model (Equation 4.1) was established, and the results showed that the social mirroring of rain patterns extracted from Twitter data is correlated with the population, income, and number of trips that arrive in the spatial units (Table 10). Only the number of trips solely explains 60% of the variability of the social mirroring of rain patterns, which implies human mobility data has more advantages to understanding social media activity related to the rain phenomenon than the other variables.

Nevertheless, the case study has several limitations that need to be properly discussed. First, spatio-temporal patterns of Twitter activity can change according to the landmark sites, point of interest, periods of the day, and on the days of the week. For example, the highest rates of georeferenced tweets are more likely to occur every day between 08:00 in the morning and midnight in Los Angeles (LI; GOODCHILD; XU, 2013). Hence, categorical variables to divide the Twitter activity in slices or spatial econometric models (i.e., using panel data rather than cross-sectional data) should be further investigated in future works.

Second, the aggregation of point-based data (e.g., tweets and rainfall rates from rain gauges) inevitably lead to the MAUP (FOTHERINGHAM; WONG, 1991). Thus, the areal unit of 30 km² should be considered only within this case study, i.e., the areal unit of analysis can change in function of different cities, data, and periods. Nonetheless, an appropriate areal unit should be chosen using systematic methods based on clear criteria that make sense to the case study, such as those presented in Section 2.1 (Chapter 2). Although the multicriteria optimization framework was used for selecting the areal unit of 30 km², as described in Section 4.2, the conceptual specification model (Section 4.3) could also have been used. However, two considerations should be taken into account: (i) while the conceptual specification model uses extrinsic factors

(i.e., demographic, socioeconomic and human mobility data), the multicriteria optimization framework can use only intrinsic features of Twitter data (e.g., global and local measures of spatial association), and (ii) while the OLS regression optimizes a single criterion (or loss function) based on a number of explanatory variables that correspond to different spatial processes (i.e., population, income, and the number of trips that arrive into the spatial units), the multicriteria optimization framework is more flexible; and may or not consider external and underlying processes.

Third, the intensity of the social mirroring of rain patterns over the spatial units was obtained by assuming synchronism between both times series of rain-related geotagged tweets and rainfall rates (lag-unit = 0). Nonetheless, such synchronism is not completely homogeneous throughout the spatial units, and different lags should be evaluated. For example, the highest cross-correlation value in a spatial unit occurs at lag-unit = 0 (synchronized), while in other spatial units could occur at lag-unit = -1 (unsynchronized). However, I contend that synchronism still makes sense in this case study, since the cross-correlation was performed for the entire period and by using a “coarse” 1-day temporal scale. In other words, two times series tend to be synchronized as the temporal scale increases; this is explained by the temporal effect of aggregation (CHENG; ADEPEJU, 2014).

Finally, the OLS regression is based on certain assumptions, among them the spatial randomness of residuals. Whenever there is statistically significant spatial autocorrelation of residuals, spatial models such as Spatial Autoregressive Regression (SAR) and Spatial Error Regression (SER) should be applied. This spatial autocorrelation indicates potential spatial effects, which yields misspecification of the OLS regression model.

4.6 Final remarks

Understanding the mechanism behind the spatial distribution of social media activity is one of the major issues to be addressed for monitoring and predicting physical phenomena using this data source; including natural phenomena such as heavy rain and flooding. This knowledge enables us to know in advance where and in which circumstances social media data can be used as a proxy data source for such phenomena. Most studies have explained social media activity through demography and socioeconomic factors; that have only depicted the population dynamic of social media by the perspective from where people live. In this chapter, I argued that the view of how people move around the city is more realistic to explain the social media activity, especially those around a topic; since they tend to present a higher spatial heterogeneity level. To support my argument, an OLS regression model was established for explaining the social media activity related to the rain phenomenon using an independent variable derived from the origin-destination matrix.

I expect this chapter leads to further works about how the mobility data can support

the explanation of spatio-temporal, thematic patterns from social media, especially in modern urban centers where the size of social media population is significantly large and social media activity can be used to infer the intra- and inter-zonal commuting of people. Future works should investigate other variables (28% of the spatial variability was not explained) and if intensities of O&D matrix could or not means spatial weights in spatial models.

A KEYWORD SENSITIVITY ANALYSIS

Keyword-based filtering is perhaps the simplest, fastest, and most common approach to retrieve thematic messages in social media analytics (STEIGER; ALBUQUERQUE; ZIPF, 2015). In this approach, a set of terms related to a topic of interest is used to determine what messages will be retrieved over a social media's Application Programming Interface (API) or filtered from a general social media dataset. Most of the studies that explore spatio-temporal patterns of natural phenomena from social media data have used keyword-based filtering to support data extraction related to forest fire (SPINSANTI; OSTERMANN, 2013), earthquakes (EARLE; BOWDEN; GUY, 2012; SAKAKI; OKAZAKI; MATSUO, 2010), hurricanes (KRYVASHEYEU *et al.*, 2016), and floods (ROSSER; LEIBOVICI; JACKSON, 2017; ALBUQUERQUE *et al.*, 2015; SMITH *et al.*, 2017).

Even widely used as an information retrieval approach to support the study of spatio-temporal, thematic patterns from social media, keyword-based filtering is prone to bias and semantic ambiguities. In general, people express themselves differently on social media platforms – they utilize different terms to communicate their experiences, feelings, and observations. This is also true to report natural phenomena who they observed or heard about. Thus, social media messages are taken of individual perceptions. In addition, the terms used to compose the messages depends on the idiosyncratic context where people are in (CURINI; IACUS; CANOVA, 2015). For example, the term “drizzle” can refer to the city of Sao Paulo (known as “the land of drizzle” by Brazilians), while the term “Sao Paulo” can refer to the Brazilian soccer team instead. A semantic analysis can reveal the true meaning of the term; nonetheless, it is a hard computational task because of the inherent variations, misspelling, and typos present in social media messages.

In light of this, ambiguous terms can lead to false-positive messages, i.e., those messages that contain the term but are not related to study context; that hereafter will be referred as noise. At the same time, ambiguous terms also lead to true-positive messages; that hereafter will be referred as signal. But, “how can the appropriate keywords for reducing the noise be chosen?” This question can be answered through a keyword sensitivity analysis of keywords. This analysis

is capable of indicating the most appropriate keywords to be used in the keyword-based filtering approach.

This chapter conducts a case study to measure the signal and noise rates of a set of flood- rain-related keywords in the city of Sao Paulo, Brazil. The case study is supported by a spatio-temporal sensitivity analysis that helps the analyst to answer the following questions:

- What are the signal and noise rates that a given keyword related to the phenomenon of interest carries in?
- Does keyword signal/noise change over space and time?

The remainder of the chapter is structured as follows: [Section 5.1](#) describes the methodology employed to assess the signal and noise rate of the keywords, as well as the case study; [Section 5.2](#) presents the results; [Section 5.3](#) conducts a discussion on the findings; finally, [Section 5.4](#) presents the final remarks.

5.1 Case study and methodology

Study area

The city of Sao Paulo was selected as the study area because it is the largest city in Brazil, with a population of approximately 12 million people ([IBGE, 2010](#)). People who live in Sao Paulo are connected to each other and social media platforms like Twitter is part of their daily routine. In addition, they share information about their everyday lives, including posts involving rain events and flooding. The city is divided in 96 districts ([Figure 28](#)), which were used as spatial units of observation for the spatio-temporal sensitivity analysis of flood- rain-related keywords.

Twitter data and keywords

A dataset consisting of tweets was retrieved through the publicly available Twitter Streaming API. Two bounding-box filters covering the city of Sao Paulo were applied (location filter), one north (-46.95, -23.62, -46.28, -23.33) and one south (-46.95, -23.91, -46.28, -23.62). A total of 11,848,923 million public tweets were retrieved from 7 November 2016 to 28 February 2017, where only 891,367 thousand tweets are georeferenced (7.52%).

After retrieving the public tweets, I selected the geotagged ones based on a set of keywords ([Chart 1](#)) – by using a substring-searching approach rather than exact word matching, i.e., the keyword chuva (rain) also matches with diadechuva (a rainy day). I filtered the ones that contained at least one of the keywords and aggregated them by districts ([Figure 28](#)).

Chart 1 – Keywords in Brazilian-Portuguese with their English meaning in parentheses. The keywords were chosen based on previous work (ANDRADE *et al.*, 2017) and a preliminary analysis of the Twitter messages. Similar terms as “chuva” (rain) and “chuvaaa” (rainn) were aggregated. Keywords with grammar mistakes were taken into account as long as the frequency was equal or greater than 10.

alagamento (flood), alagado (flooded), alagada (flooded), alagando (it's flooding), alagou (flooded), alagar (to flood), chove (rain), chova (rain), chovia (had been rained), chuva (rain), chuvarada (rain), chuvosa (rain), chuvoso (rainy), chuvona (heavy rain), chuvinha (drizzle), chuvisco (drizzle), chuvendo (it's raining), dilúvio (heavy rain), garoa (drizzle), inundação (flood), inundada (flooded), inundado (flooded), inundar (to flood), inundam (flood), inundou (flooded), temporal (storm), temporais (storms)

Figure 28 – Choropleth map of the number of filtered tweets per district of the city of Sao Paulo.

Source: Elaborated by the author.

Keyword sensitivity analysis

The keyword sensitivity analysis consisted of two steps: (i) labeling the 5,408 filtered tweets as on-topic and off-topic, and (ii) extracting the signal and noise time series of the keywords.

First, five raters manually labeled 3,964 geotagged tweets as on-topic (related to local rain or flood), and 1,444 geotagged tweets as off-topic (not related to local rain or flood). In the following, the degree of agreement among raters was determined by means of the Krippendorff's alpha coefficient, a statistical measure of the degree of agreement among two or more raters (KRIPPENDORFF, 2004). A value equal to 0.72 was obtained, which indicates a moderate agreement among raters. A coefficient equals to 0 (zero) indicates an absence of agreement and 1 (one) a perfect agreement.

Second, for each district and keyword, two time series were extracted, called signal and noise time series. These time series correspond to the frequency of on-topic and off-topic geotagged tweets, respectively. For this, 5 temporal scales of aggregation were used: (i) 30 min, (ii) hour, (iii) 12 hours, (iv) day, and (v) week. After that, a third time series was built with the ratio between the frequency of on-topic/off-topic (signal/noise) geotagged tweets and the total frequency of geotagged tweets. Finally, the third time series, named relative frequency, was analyzed over temporal scales and across the districts. The main idea was to evaluate the difference between the time series of the signal and noise for each keyword.

5.2 Results

The results show that the keyword noise changes over time and space, leading to depreciation or an increase in the keyword signal (Figure 29). Moreover, the signal tends to increase (appear) for large temporal scales (e.g., weeks) and decrease (disappear) for small temporal scales (e.g., minutes). In addition, there is a spatial dependence of the keyword signal across the districts, i.e., the signal and noise are usually more similar in near districts than distant ones. For example, the Sé district (Figure 29 (a)) is similar to the Barra Funda district (Figure 29 (b)), whereas the Cidade Dutra district (Figure 29 (c)) is different from both districts (Figure 29 (a) and Figure 29 (b)). However, the distance sometimes does not influence the similarity among the districts. For example, the Sé and Itaquera districts (Figure 29 (a) and Figure 29 (d)) are similar and far away from each other. That means that the amount of geotagged tweets posted within the districts (Figure 28) does not explain totally the signal of the keywords. Other variables, such as the interconnection areas of the underground railway system and economic factors, could also describe the signal (see Chapter 4).

When the keywords are examined, we can see that some of them do not vary over time, such as “chuvinha” (raining a little), “chuvosa” (rainy) and “inundação” (flood), i.e., they do not often vary from signal to noise or vice-versa. On the other hand, some keywords reveal greater

noise in short time intervals, such as the keyword “chuva” (rain).

Furthermore, some keywords have the potential to compose a (good) signal; however, they create noise. This could be explained by the fact that some words have a special association with local language/culture or atypical events. The keyword “garoa” (drizzle), for instance, might be related to the drizzle phenomenon; however, most messages refer to the codename of the city of Sao Paulo (“the land of drizzle”). Other interesting example occurred during the concert of the rock band Guns and Roses at the Allianz Park in the Barra Funda district. On 12 November 2016, there was a frequency peak of the keyword “chuva” (rain) when the band played one of their most famous songs, “November Rain”. Messages like “chuva de novembro” (november rain) and “a chuva veio antes pra colocar todo mundo no clima da November Rain!...” (the rain came before to put everyone in the November Rain mood! ...) were reported by people who were attending the concert.

Figure 29 – Hovmöller-based diagram depicting the signal (1.0) and noise (-1.0) of the keywords over five temporal scales of aggregation and across four highlighted districts (see Figure 28). The x and y axes show the temporal scales of aggregation and the keywords, respectively. The blue color represents the signal intensity, whereas the red color represents the noise intensity. White color represents no data. The signal and noise were measured as the fraction between on-topic and off-topic geotagged tweets and all the geotagged tweets posted within the district (relative frequency) and, later, rescaled to [-1, 1].

Note – Continued on next page.

Note – Continued on next page.

Source: Elaborated by the author.

The underlying problem behind using keywords is their reproducibility from one area to another or at in same area over time. [Rzeszewski \(2018\)](#) refers to this behaviour as a change of the perception of the physical space. As shown in [Figure 29 \(a-d\)](#), the behavior changes in terms of time and space. Hence, keywords should be selected with caution, considering local issues, such as language/culture, and, especially, atypical events.

5.3 Discussion

The results evidence that keywords on social media analysis are sensitive to time and space ([Section 5.2](#)). At first glance, all predefined keywords ([Chart 1](#)) had the potential to filter rain- and flood-related messages from social media data; however, some keywords are noisy and may filter false-positive messages when used in a keyword-based filtering approach. For example, people usually post messages with the keyword “garoa” (drizzle) as reference to the city of Sao Paulo (known as “the land of drizzle”), which could lead to a noisy social media dataset for hydrometeorological analysis. Therefore, the choice of keywords can influence the performance of the keyword-based filtering approach; which has been a useful tool to reduce the amount of data in social media analytics.

In this chapter, I argue that the choice of the most appropriate keywords or terms to filter messages in social media analytics should be supported by a sensitivity analysis. This analysis allows us to identify, from an initial set of keywords, which ones can filter more true-positive messages rather than false-positive messages. The case study carried in the city of Sao Paulo showed that the keywords “chove”, “chuva”, “chuvinha”, and “chuvisco” produce good signals across the districts and at different temporal scales of aggregation.

Nevertheless, there are some shortcomings in this case study that could not be addressed. The limitation of the Twitter Streaming API for stratified sampling introduces biases in the inter-district spatial analysis. Clearly, the number of filtered tweets change throughout the districts and the results do not take it into account. Ideally, the sampling method should be controlled in spatial analysis.

Another limitation refers to the establishment of an acceptable noise rate for keyword selection, i.e., a threshold. Further investigation should be done to determine such a threshold, including periods such as rainy and non-rainy seasons. However, I believe this threshold must be locality-driven and, thus, should not be generalized to different cities.

5.4 Final remarks

The keyword-based filtering approach aims at assisting the data reduction in social media studies around a topic. A set of keywords is usually chosen based on personal knowledge about the topic in hand. However, the way people report real-world phenomena is influenced by idiosyncratic and subjective local contexts, leading some keywords to the retrieval of a large amount of false-positive messages. In this chapter, I argue that a sensitivity analysis should support the choice of the most appropriate keywords to filter messages in social media analytics. This analysis allows us to identify, from an initial set of keywords, which ones can filter more true-positive messages rather than false-positive ones.

Future work should further extend this keyword sensitivity analysis by incorporating other cities and terms. Finally, skip-gram models (e.g., word embedding) and Latent Dirichlet Allocation could be coupled with keyword-based filter approach to address the ambiguity problem of terms in social media analysis.

CONCLUSIONS

This doctoral research has demonstrated that Twitter data can be a valuable and cost-effective data source for supporting flood risk management through a proper mining of rain patterns. Rain patterns are fundamental inputs in many rainfall-runoff models for flood modeling and forecasting, as well as for early warning systems of extreme weather. Therefore, the establishment of social signals that reflect rain phenomenon has opened up a new way of exploring and improving solutions for flood risk management. The challenges and issues that support this practice and accurately mirror rain patterns from Twitter data were properly addressed throughout the chapters.

[Chapter 2](#) discussed the effects of MAUP on rain patterns extracted from Twitter data and the dependence of an appropriate areal unit for mirroring the rain phenomenon accurately. Areal unit selection approaches, nonetheless, share the assumption the most suitable areal unit can be determined on the basis of a single optimization criterion (or loss function) ([LEE *et al.*, 2016](#); [OPENSHAW, 1977](#)). However, this research has demonstrated multiple and conflicting criteria might determine the choice of a suitable areal unit of analysis, and a novel and innovative method for the selection of a more appropriate areal unit of aggregation in social media analysis was proposed and evaluated. The method relies on Pareto optimality and assesses a set of candidate areal units of aggregation vis-à-vis a number of optimization criteria. Moreover, it can be adapted to include any number of criteria that may be appropriate for different spatial datasets and applications.

[Chapter 3](#) examined the temporal validity of rain patterns extracted from Twitter data. The problem refers to synchronism (or lack thereof) between the temporal units of the aggregated patterns from social media data and a given phenomenon of interest. A temporal approach was proposed and evaluated for the determination of the time interval that validates the rain patterns extracted from Twitter data for nowcasting or forecasting. It relies on the cross-correlation technique and enables analyses of similarity of two times-series through temporal units and their lag-time.

Chapter 4 addressed the characterization of the spatial units of analysis where Twitter activity accurately mirrors rain patterns. Traditionally, the spatial distribution of Twitter activity is associated with census data, such as population, income, and well-educated variables (SLOAN, 2017; LONGLEY; ADNAN, 2016; XIAO; HUANG; WU, 2015; LI; GOODCHILD; XU, 2013). Nonetheless, a conceptual specification model was established to show human mobility data source is one of the main components behind the spatial distribution of Twitter activity, since it describes the human density in cities through the movement of people between areas. Therefore, human mobility data explain the accuracy of the rain patterns extracted from Twitter data more than census data.

Last, but not least, Chapter 5 reported on a twofold contribution. First, a sensitivity analysis identified, from an initial set of rain- and flood-related keywords, the ones that improve the performance of the keyword-based filtering approach in social media analysis of rain patterns. Performance here refers to the number of true-positive messages retrieved through social media Streaming APIs, or directly on social media data. Second, the performance of the keyword-based filtering approach that uses such keywords changes in time and space, which requires attention to its application.

6.1 Contributions

In summary, the following contributions have been made:

- (i) a systematic method for the selection of an optimal areal unit in social media analytics;
- (ii) a temporal approach for the evaluation of the temporal validity of social media activity related to a given natural phenomenon of interest;
- (iii) establishment of a conceptual specification model that characterizes spatial units where the social media activity accurately mirrors a given phenomenon of interest; and
- (iv) a sensitivity analysis of the spatio-temporal patterns of keywords related to a given phenomenon of interest.

Such contributions comprise a three dimensional-based approach in terms of space, time, and content, for studying several types of continuous physical phenomena from social media data. The way how such dimensions were analyzed (i.e., the methodological approach) provides advances in social media analytics. Furthermore, the methodology can be transferred to assist temporal and spatial analysis of other types of human activities and physical phenomenon, providing advances in the field of GIScience.

6.2 Limitations

Nevertheless, some limitations should be considered.

- **Case studies**

The case studies conducted are restricted to Twitter data in the city of Sao Paulo, Brazil. Therefore, the findings are bounded by the characteristics of the Twitter platform and the city's geography, and should not be generalized or extended to other cities and social media platforms without a reproducibility of the case studies. Different findings might be made changing cities and social media platforms. For example, the 'optimal' areal units of 30 km² and 50 km² might be 'non-optimal' to other cities, even those sharing similar geographic features to Sao Paulo city (Chapter 2). In Chapter 3, the range of lag-units between the times series of rainfall measurements and frequency of rain-related geotagged tweets might be greater than -1 to +1. And mobility data might not suitable explain the social media activity in small cities (Chapter 4).

- **Data collection process**

There is a potential for some biases in the case studies owing to unrepresentativeness regions/data and missing data. The limitations of the Twitter API Streaming (e.g., the rate limit when the stream returns more than 1% of the all tweets being tweeted at a given moment) and failures in the client application for retrieving messages from Twitter (e.g., connection drops) (ANDRADE *et al.*, 2018b) might introduce obstacles in finding more meaningful relationships between the data sources (i.e., Twitter, weather radar, and rain gauges). However, it is unlikely that such problems would completely reverse the statistical relationships observed between Twitter data related to the rain phenomenon and rainfall data provided from the physical sensors. Moreover, the geotagged rain-related tweets were retrieved only based the available location (i.e., coordinates). Some tweets, on the other hand, refer to phenomenon location by hashtags or in the message content. Therefore, strategies might be used to fetch those references towards geolocating tweets and, thus, increase the sample size.

- **Applications to flood risk management**

The findings are limited to proofs of concepts or post-hoc analyses that use real data. Some issues, such as the impact of this new data source on the theory of hydrological models for flood risk management, should be investigated regarding the use of Twitter data in applications of rainfall-runoff models for monitoring and forecasting. Furthermore, integration issues of rain patterns extracted from Twitter data and physical sensors have been marginally addressed and are limited to correlation analyses.

- **Frontier with the social sciences**

There is no reason to believe that the limitations are strictly linked to technical and quantitative aspects, i.e., which there are no limitations related to other non-technical areas, such as social sciences. Social media activity depends on people and how they interact with each other and with the place who they live. Such issues were not addressed in this doctoral research, but they might clarify several aspects of the production of social media data and the influences of idiosyncratic and subjective local contexts (e.g., the lack of representativeness of certain social groups).

- **Framing of the contributions**

The limitations of the four contributions (the above-mentioned items (i)–(iv)) were individually discussed in each Chapter. Readers are invited to reread [Section 2.5](#) for limitations of the multicriteria optimization framework, [Section 3.5](#) for limitations of the temporal approach, [Section 4.5](#) for limitations of the conceptual specification model, and [Section 5.3](#) for limitations of the keywords sensitivity analysis.

6.3 Recommendations for future work

The conclusions drawn and the knowledge gained from this doctoral thesis enable a wide range of further research. Therein, the following recommendations have been identified as particularly valuable:

- **Cost-benefit analysis**

The application of rainfall-runoff models for monitoring and forecasting using social media data should be investigated through a cost-benefit analysis, since the massive amount of data provided by different social media platforms can lead to updates in infrastructure and data management systems – including new components of collection, storage, retrieval, and preprocessing of unstructured data. Therefore, the investment's gains from the use of social media data as a supplementary data source should compare favorably in relation to cost.

- **Classification of social media messages**

The retrieval of relevant and meaningful messages from social media is not a straightforward task. The heterogeneous, ambiguous, and misspelled content of messages hampers their automatic classification, and requires a preprocessing and application of robust models that very often hinder the speed of the use combined with conventional data sources. Furthermore, the classification task requires a well-defined set of categories that heavily depends on the phenomenon, geographic context, and purpose. Moreover, it is still necessary to rely the annotations and tags that humans attach to social media messages.

- **Prioritization of social media messages**

Most studies on social media analytics and situational awareness are based on Tobler's first law of geography, which assumes social media messages nearest to affected areas (or phenomenon of interest) are more relevant than more distant ones. However, such studies cannot measure the amount of information conveyed by these messages, which might be useful for prioritizing and showing only ones most likely to assist decision makers. A content-driven prioritization approach in combination with geographical distance could be a powerful tool for assisting decision-makers during crises and natural disasters. This topic is on our current research agenda, although the results achieved so far are incipient (ANDRADE *et al.*, 2018c).

Other minor recommendations for future works have also been individually mentioned for each contribution (i.e., for the above-mentioned items (i)–(iv)). Readers are invited to reread Sections 2.5, 3.5, 4.5, and 5.3.

Finally, in an ideal world, the inclusion of a new data source for supporting disaster risk management should match its design; however, the real world requires adaptations, preprocessing, processing, and reprocessing. This doctoral thesis did not aim to present designs and rules, but guide principles for the mining of natural phenomena patterns from social media data.

6.4 Data and codes availability statement

In order to ensure reproducibility and replicability of this doctoral thesis, the data and codes that support the contributions and findings are available at <<https://doi.org/10.6084/m9.figshare.11369649>> and <<http://wrap.warwick.ac.uk/87173/>>.

BIBLIOGRAPHY

AHRENS, B. Distance in spatial interpolation of daily rain gauge data. **Hydrology and Earth System Sciences**, v. 10, n. 2, p. 197–208, 2006. Citation on page 60.

ALBUQUERQUE, J. P. de; HERFORT, B.; BRENNING, A.; ZIPF, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. **International Journal of Geographical Information Science**, Taylor & Francis, v. 29, n. 4, p. 667–689, 2015. Citations on pages 27, 42, 58, and 87.

ANDRADE, S. C. de; DEGROSSI, L. C.; ESTRADA, C. R.; DELBEM, A. C. B.; ALBUQUERQUE, J. P. de. Does keyword noise change over space and time? A case study of social media messages. In: **XIX Brazilian Symposium on Geoinformatics – GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, 2018**. [S.l.: s.n.], 2018. p. 116–121. Citations on pages 31, 42, and 75.

ANDRADE, S. C. de; QUEIROZ, L. H.; DELBEM, A. C. B.; ALBUQUERQUE, J. P. de. An architectural model for retrieving data from Social Media APIs: an instantiation for consuming the Twitter Stream API. **Poster presented at the XIX Brazilian Symposium on Geoinformatics – GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, 2018**. 2018. Citation on page 97.

ANDRADE, S. C. de; RESTREPO-ESTRADA, C.; COSTA, T. A. G. da; UEYAMA, J.; DELBEM, A. C. B.; ALBUQUERQUE, J. P. de. Situational awareness in social media: lessons learned using information entropy in flood risk management. In: **Workshop @NUVEM**. [S.l.]: UFABC, 2018. Citation on page 99.

ANDRADE, S. C. de; RESTREPO-ESTRADA, C.; DELBEM, A. C. B.; MENDIONDO, E. M.; ALBUQUERQUE, J. P. de. Mining rainfall spatio-temporal patterns in twitter: A temporal approach. In: BREGT, A.; SARJAKOSKI, T.; LAMMEREN, R. van; RIP, F. (Ed.). **Societal Geo-innovation**. Cham: Springer International Publishing, 2017. p. 19–37. ISBN 978-3-319-56759-4. Citations on pages 17, 29, 31, 49, and 89.

ANDRADE, S. C. de; RESTREPO-ESTRADA, C.; NUNES, L. H.; DELBEM, A. C. B.; RODRIGUEZ, C. A. M.; ESTRELLA, J. C.; ALBUQUERQUE, J. P. de. A multicriteria optimization framework for the definition of the spatial granularity of urban social media analytics. **International Journal of Geographical Information Science**, Taylor & Francis, v. 0, n. 0, 2020. Citations on pages 30 and 76.

ANSELIN, L. **Spatial Econometrics: Methods and Models**. Dordrecht: Kluwer Academic Publishers, 1988. Citations on pages 32, 72, 76, 109, 110, 111, and 112.

_____. Local Indicators of Spatial Association–LISA. **Geographical Analysis**, v. 27, n. 2, p. 93–115, 1995. Citations on pages 36, 37, 49, and 112.

ARAGÓN-DURAND, F. **Inundaciones en zonas urbanas de cuencas en América Latina**. Lima, Perú: Soluciones Prácticas, Lima, 2014. Available: <<https://infoinundaciones.com/>>

[recursos/item/inundaciones-en-zonas-urbanas-de-cuencas-en-america-latina>](#). Citation on page 25.

ARTHUR, R.; BOULTON, C. A.; SHOTTON, H.; WILLIAMS, H. T. P. Social sensing of floods in the UK. **PLOS ONE**, Public Library of Science, v. 13, n. 1, p. 1–18, 01 2018. Citations on pages 27 and 36.

ASSIS, L. F. G.; HERFORT, B.; STEIGER, E.; HORITA, F. E. A.; ALBUQUERQUE, J. P. de. Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods. In: **Proceedings of the XVI Brazilian Symposium on GeoInformatics**. [S.l.: s.n.], 2015. p. 26–37. Citations on pages 57 and 69.

BACCHI, B.; KOTTEGODA, N. T. Identification and calibration of spatial correlation patterns of rainfall. **Journal of Hydrology**, v. 165, n. 1, p. 311–348, 1995. ISSN 0022-1694. Citation on page 60.

BARTIER, P. M.; KELLER, C. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (idw). **Computers & Geosciences**, v. 22, n. 7, p. 795–799, 1996. ISSN 0098-3004. Citation on page 60.

BATTAN, L. J. Radar observation of the atmosphere. L. J. Battan (The University of Chicago Press) 1973. PP X, 324; 125 figures, 21 tables. £7·15. **Quarterly Journal of the Royal Meteorological Society**, v. 99, n. 422, 1973. Citation on page 43.

BATTY, M. **The New Science of Cities**. [S.l.]: MIT Press, 2013. (The MIT Press). ISBN 978-0262534567. Citations on pages 26 and 32.

BLÖSCHL, G.; SIVAPALAN, M. Scale issues in hydrological modelling: A review. **Hydrological Processes**, v. 9, n. 3–4, p. 251–290, 1995. Citations on pages 50 and 54.

CARLEY, K. M.; MALIK, M.; LANDWEHR, P. M.; PFEFFER, J.; KOWALCHUCK, M. Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. **Safety Science**, v. 90, p. 48–61, 2016. ISSN 0925-7535. Building Community Resilience to Global Hazards: A Sociotechnical Approach. Citation on page 53.

CARR, D. B.; OLSEN, A. R.; WHITE, D. Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data. **Cartography and Geographic Information Systems**, Taylor & Francis, v. 19, n. 4, p. 228–236, 1992. Citation on page 41.

CHENG, T.; ADEPEJU, M. Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection. **PLOS ONE**, Public Library of Science, v. 9, n. 6, p. 1–10, 06 2014. Citations on pages 28, 49, 65, and 84.

CHENG, Z.; CAVERLEE, J.; LEE, K.; SUI, D. Exploring millions of footprints in location sharing services. In: . [S.l.: s.n.], 2011. p. 81–88. Citation on page 71.

COLLETTE, Y.; SIARRY, P. **Multiobjective Optimization**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. 522 p. (Decision Engineering). ISBN 978-3-642-07283-3. Citation on page 39.

CRED EM-DAT. **The OFDA/CRED - International Disaster Database**. 2015. www.emdat.be Université catholique de Louvain Brussels - Belgium. Citation on page 25.

CURINI, L.; IACUS, S.; CANOVA, L. Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case. **Social Indicators Research**, v. 121, n. 2, p. 525–542, 2015. ISSN 1573-0921. Citation on page [87](#).

DARK, S. J.; BRAM, D. The modifiable areal unit problem (MAUP) in physical geography. **Progress in Physical Geography: Earth and Environment**, v. 31, n. 5, p. 471–479, 2007. Citations on pages [28](#), [33](#), [34](#), [35](#), and [36](#).

EARLE, P.; BOWDEN, D.; GUY, M. Twitter earthquake detection: earthquake monitoring in a social world. **Annals of Geophysics**, v. 54, n. 6, p. 708–715, 2012. Citations on pages [27](#), [54](#), and [87](#).

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 1979. ISSN 00905364. Citation on page [40](#).

ELWOOD, S.; LESZCZYNSKI, A. Privacy, reconsidered: New representations, data practices, and the geoweb. **Geoforum**, v. 42, n. 1, p. 6–15, 2011. ISSN 0016-7185. Citation on page [27](#).

FAHRIG, L.; BAUDRY, J.; BROTONS, L.; BUREL, F. G.; CRIST, T. O.; FULLER, R. J.; SIRAMI, C.; SIRIWARDENA, G. M.; MARTIN, J.-L. Functional landscape heterogeneity and animal biodiversity in agricultural landscapes. **Ecology Letters**, v. 14, n. 2, p. 101–112, 2011. Citations on pages [112](#) and [113](#).

FONSECA, V. Grunert da; FONSECA, C. M.; HALL, A. O. Inferential Performance Assessment of Stochastic Optimisers and the Attainment Function. In: ZITZLER, E.; THIELE, L.; DEB, K.; COELLO, C. A. C.; CORNE, D. (Ed.). **Evolutionary Multi-Criterion Optimization**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 213–225. ISBN 978-3-540-44719-1. Citation on page [40](#).

FOTHERINGHAM, A. S.; WONG, D. W. S. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. **Environment and Planning A: Economy and Space**, v. 23, n. 7, p. 1025–1044, 1991. Citations on pages [36](#) and [83](#).

FOTHERINGHAM, S. Scale-independent spatial analysis. In: GOODCHILD, M.; GOPAL, S. (Ed.). **The accuracy of spatial databases**. [S.l.]: Taylor & Francis, 1989. p. 144–149. Citations on pages [28](#), [33](#), [34](#), and [35](#).

GETIS, A. Reflections on spatial autocorrelation. **Regional Science and Urban Economics**, v. 37, n. 4, p. 491–496, 2007. ISSN 0166-0462. Citation on page [110](#).

GOODCHILD, M. F. Geographical information science. **International Journal of Geographical Information Systems**, Taylor & Francis, v. 6, n. 1, p. 31–45, 1992. Citations on pages [25](#), [32](#), and [109](#).

_____. Twenty years of progress: Giscience in 2010. **Journal Spatial Information Science**, , v. 1, n. , p. 3–20, 2010. Citations on pages [32](#) and [109](#).

GOTWAY, C. A.; YOUNG, L. J. Combining incompatible spatial data. **Journal of the American Statistical Association**, Taylor & Francis, v. 97, n. 458, p. 632–648, 2002. Citation on page [77](#).

GRECO, S.; EHRGOTT, M.; FIGUEIRA, J. R. **Multiple Criteria Decision Analysis: State of the Art Surveys**. New York, NY: Springer, New York, NY, 2016. (International Series in Operations Research & Management Science, v. 233). ISBN 978-1-4939-3093-7. Citations on pages 38 and 39.

HAINING, R. **Spatial Data Analysis: Theory and Practice**. [S.l.]: Cambridge University Press, 2003. Citations on pages 32, 72, and 109.

HUANG, Q.; XIAO, Y. Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery. **ISPRS International Journal of Geo-Information**, v. 4, n. 3, p. 1549–1568, 2015. ISSN 2220-9964. Citations on pages 42 and 58.

HUANG, X.; WANG, C.; LI, Z. A near real-time flood-mapping approach by integrating social media and post-event satellite imagery. **Annals of GIS**, Taylor & Francis, v. 24, n. 2, p. 113–123, 2018. Citation on page 27.

IBGE. **Censo Demográfico 2010**. Rio de Janeiro: Brazilian Institute of Geography and Statistics, 2010. Citations on pages 41, 74, and 88.

IMRAN, M.; CASTILLO, C.; DIAZ, F.; VIEWEG, S. Processing social media messages in mass emergency: A survey. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 47, n. 4, p. 67:1–67:38, Jun. 2015. ISSN 0360-0300. Citations on pages 27, 53, and 73.

JIANG, B. Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. **GeoJournal**, v. 80, n. 1, p. 1–13, Feb 2015. ISSN 1572-9893. Citation on page 49.

JIANG, B.; BRANDT, S. A. A Fractal Perspective on Scale in Geography. **ISPRS International Journal of Geo-Information**, v. 5, n. 6, p. 95, 2016. ISSN 2220-9964. Citations on pages 35 and 49.

JIANG, B.; MIAO, Y. The Evolution of Natural Cities from the Perspective of Location-Based Social Media. **The Professional Geographer**, Routledge, v. 67, n. 2, p. 295–306, 2015. Citations on pages 33, 36, and 72.

JURDAK, R.; ZHAO, K.; LIU, J.; ABOUJAOUDE, M.; CAMERON, M.; NEWTH, D. Understanding human mobility from twitter. **PLOS ONE**, Public Library of Science, v. 10, n. 7, p. 1–16, 07 2015. Citations on pages 71, 76, and 83.

KOUNADI, O.; RESCH, B.; PETUTSCHNIG, A. Privacy threats and protection recommendations for the use of geosocial network data in research. **Social Sciences**, v. 7, n. 10, 2018. ISSN 2076-0760. Citation on page 27.

KRIPPENDORFF, K. Reliability in content analysis: Some common misconceptions and recommendations. **Human Communications Research**, v. 30, n. 3, p. 411–433, 2004. Citation on page 90.

KRUSKAL, J. B.; LIBERMAN, M. The symmetric time-warping problem: from continuous to discrete. In: SANKOFF, D.; KRUSKAL, J. B. (Ed.). **Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison**. [S.l.]: CSLI Publications, Stanford, CA 94305, 1999. chap. 4. Citation on page 56.

KRYVASHEYEU, Y.; CHEN, H.; OBRADOVICH, N.; MORO, E.; HENTENRYCK, P. V.; FOWLER, J.; CEBRIAN, M. Rapid assessment of disaster damage using social media activity. **Science Advances**, American Association for the Advancement of Science, v. 2, n. 3, 2016. Citations on pages [27](#) and [87](#).

LANDWEHR, P. M.; CARLEY, K. M. Social media in disaster relief. In: _____. **Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. p. 225–257. ISBN 978-3-642-40837-3. Citation on page [27](#).

LEE, Y.; KWON, P.; YU, K.; PARK, W. Method for Determining Appropriate Clustering Criteria of Location-Sensing Data. **ISPRS International Journal of Geo-Information**, v. 5, n. 9, 2016. ISSN 2220-9964. Citations on pages [33](#), [35](#), [36](#), [48](#), and [95](#).

LI, L.; GOODCHILD, M. F.; XU, B. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. **Cartography and Geographic Information Science**, Taylor & Francis, v. 40, n. 2, p. 61–77, 2013. Citations on pages [29](#), [71](#), [73](#), [83](#), and [96](#).

LI, Z.; WANG, C.; EMRICH, C. T.; GUO, D. A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. **Cartography and Geographic Information Science**, Taylor & Francis, v. 45, n. 2, p. 97–110, 2018. Citation on page [27](#).

LONGLEY, P. A.; ADNAN, M. Geo-temporal Twitter demographics. **International Journal of Geographical Information Science**, Taylor & Francis, v. 30, n. 2, p. 369–389, 2016. Citations on pages [27](#), [36](#), [71](#), [73](#), [83](#), and [96](#).

MAIR, A.; FARES, A. Comparison of rainfall interpolation methods in a mountainous region of a tropical island. **Journal of Hydrologic Engineering**, v. 16, n. 4, p. 371–383, 2011. Citation on page [60](#).

MANCA, M.; BORATTO, L.; ROMAN, V. M.; GALLISSÀ, O. M. i; KALTENBRUNNER, A. Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. **Online Social Networks and Media**, v. 1, p. 5–69, 2017. ISSN 2468-6964. Citation on page [27](#).

MARK, D. M. Geographic information science: Defining the field. In: DUCKHAM, M.; GOODCHILD, M. F.; WORBOYS, M. (Ed.). **Foundations of Geographic Information Science**. [S.l.]: Taylor & Francis, 2003. p. 3–18. Citations on pages [25](#) and [109](#).

MARTÍ, P.; SERRANO-ESTRADA, L.; NOLASCO-CIRUGEDA, A. Social Media data: Challenges, opportunities and limitations in urban studies. **Computers, Environment and Urban Systems**, v. 74, p. 161–174, 2019. ISSN 0198-9715. Citation on page [27](#).

MARTÍNEZ-ROJAS, M.; PARDO-FERREIRA, M. del C.; RUBIO-ROMERO, J. C. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. **International Journal of Information Management**, v. 43, p. 196–208, 2018. ISSN 0268-4012. Citation on page [27](#).

Meng, Y.; Lin, C.; Cui, W.; Yao, J. Scale selection based on Moran's I for segmentation of high resolution remotely sensed images. In: **2014 IEEE Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2014. p. 4895–4898. ISSN 2153-6996. Citation on page [36](#).

MOORE, R. J. The pdm rainfall-runoff model. **Hydrology and Earth System Sciences**, v. 11, n. 1, p. 483–499, 2007. Citation on page 27.

MORSTATTER, F.; PFEFFER, J.; LIU, H.; CARLEY, K. M. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. **CoRR**, abs/1306.5204, 2013. Citations on pages 42 and 75.

NUMMI, P. Social Media Data Analysis in Urban e-Planning. **Int. J. E-Planning Res.**, IGI Global, Hershey, PA, USA, v. 6, n. 4, p. 18–31, Oct. 2017. ISSN 2160–9918. Citations on pages 27 and 73.

OPENSHAW, S. A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. **Transactions of the Institute of British Geographers**, [Royal Geographical Society (with the Institute of British Geographers), Wiley], v. 2, n. 4, p. 459–472, 1977. ISSN 00202754, 14755661. Citations on pages 34, 35, 48, and 95.

_____. An Empirical Study of Some Zone-Design Criteria. **Environment and Planning A: Economy and Space**, v. 10, n. 7, p. 781–794, 1978. Citation on page 34.

_____. Ecological Fallacies and the Analysis of Areal Census Data. **Environment and Planning A: Economy and Space**, v. 16, n. 1, p. 17–31, 1984. Citations on pages 28, 33, and 34.

OPENSHAW, S.; RAO, L. Algorithms for Reengineering 1991 Census Geography. **Environment and Planning A: Economy and Space**, v. 27, n. 3, p. 425–446, 1995. Citation on page 35.

PATEL, N. N.; STEVENS, F. R.; HUANG, Z.; GAUGHAN, A. E.; ELYAZAR, I.; TATEM, A. J. Improving Large Area Population Mapping Using Geotweet Densities. **Transactions in GIS**, v. 21, n. 2, p. 317–331, 2017. Citation on page 27.

POORTHUIS, A. How to Draw a Neighborhood? The Potential of Big Data, Regionalization, and Community Detection for Understanding the Heterogeneous Nature of Urban Neighborhoods. **Geographical Analysis**, v. 50, n. 2, p. 182–203, 2018. Citation on page 35.

POORTHUIS, A.; ZOOK, M.; SHELTON, T.; GRAHAM, M.; STEPHENS, M. **Using Geo-tagged Digital Social Data in Geographic Research**. 2014. Pre-publication version of chapter submitted to: *Key Methods in Geography*, eds. eds. N. Clifford, S. French, M. Cope, and S. Gillespie. Citations on pages 14, 42, and 47.

RESTREPO-ESTRADA, C.; ANDRADE, S. C. de; ABE, N.; FAVA, M. C.; MENDIONDO, E. M.; ALBUQUERQUE, J. P. de. Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. **Computers & Geosciences**, v. 111, p. 148–158, 2018. ISSN 0098-3004. Citations on pages 25, 36, and 49.

RESTREPO-ESTRADA, C. E. **Uso de dados das mídias sociais no monitoramento de enchentes**. Phd Thesis (PhD Thesis) — University of Sao Paulo, The address of the publisher, 11 2018. Citations on pages 25 and 27.

RODRÍGUEZ, L.; PALANCA, J.; VAL, E. del; REBOLLO, M. Analyzing urban mobility paths based on users’ activity in social networks. **Future Generation Computer Systems**, v. 102, p. 333–346, 2020. ISSN 0167-739X. Citations on pages 71, 76, and 83.

ROOCKS, P. Computing Pareto Frontiers and Database Preferences with the rPref Package. **The R Journal**, v. 8, n. 2, p. 393–404, 2016. Citations on pages 39 and 45.

ROSSER, J. F.; LEIBOVICI, D. G.; JACKSON, M. J. Rapid flood inundation mapping using social media, remote sensing and topographic data. **Natural Hazards**, v. 87, n. 1, p. 103–120, 2017. Citations on pages 27 and 87.

RZESZEWSKI, M. Geosocial capta in geographical research – a critical analysis. **Cartography and Geographic Information Science**, Taylor & Francis, v. 45, n. 1, p. 18–30, 2018. Citations on pages 33, 72, and 93.

SAKAKI, T.; MATSUO, Y. Earthquake observation by social sensors. In: _____. **Earthq. Res. Anal. - Stat. Stud. Obs. Plan.** [S.l.]: InTech, 2012. Citation on page 69.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: **Proceedings of the 19th International Conference on World Wide Web**. [S.l.: s.n.], 2010. p. 851–860. Citations on pages 27 and 87.

SELEX. **Selex meteor manual**. [S.l.], 2015. Citations on pages 43 and 75.

SINGLETON, A. D.; SPIELMAN, S.; FOLCH, D. **Urban Analytics**. [S.l.]: SAGE, 2018. ISBN 9781473958630. Citations on pages 26, 32, and 72.

SLOAN, L. Who tweets in the united kingdom? profiling the twitter population using the british social attitudes survey 2015. **Social Media + Society**, v. 3, n. 1, p. 2056305117698981, 2017. Citations on pages 29, 71, 73, 75, 83, and 96.

SMITH, L.; LIANG, Q.; JAMES, P.; LIN, W. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. **Journal of Flood Risk Management**, v. 10, n. 3, p. 370–380, 2017. Citations on pages 27 and 87.

SPINSANTI, L.; OSTERMANN, F. Automated geographic context analysis for volunteered information. **Applied Geography**, v. 43, p. 36–44, 2013. Citations on pages 27, 69, and 87.

SPMR. **Pesquisa Origem e Destino de 2017**. São Paulo: [s.n.], 2017. Citation on page 75.

STEIGER, E.; ALBUQUERQUE, J. a. P. de; ZIPF, A. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. **Transactions in GIS**, v. 19, n. 6, p. 809–834, 2015. Citations on pages 27 and 87.

STEIGER, E.; WESTERHOLT, R.; RESCH, B.; ZIPF, A. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. **Computers, Environment and Urban Systems**, v. 54, p. 255–265, 2015. ISSN 0198-9715. Citations on pages 27 and 76.

TENKANEN, H.; MININ, E. D.; HEIKINHEIMO, V.; HAUSMANN, A.; HERBST, M.; KAJALA, L.; TOIVONEN, T. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. **Scientific Reports**, Nature, v. 7, 2017. Citation on page 36.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic Geography**, [Clark University, Wiley], v. 46, p. 234–240, 1970. ISSN 00130095, 19448287. Available: <<http://www.jstor.org/stable/143141>>. Citation on page 110.

_____. Frame independent spatial analysis. In: GOODCHILD, M.; GOPAL, S. (Ed.). **The accuracy of spatial databases**. [S.l.]: Taylor & Francis, 1989. p. 75–79. Citation on page 35.

UNISDR, U. N. I. S. for D. R. **2015 Global Assessment Report on Disaster Risk Reduction**. [s.n.], 2015. 312 p. Available: <<https://www.un-ilibrary.org/content/publication/919076d9-en>>. Citation on page 25.

WANG, Z.; YE, X.; LEE, J.; CHANG, X.; LIU, H.; LI, Q. A spatial econometric modeling of online social interactions using microblogs. **Computers, Environment and Urban Systems**, v. 70, p. 53–58, 2018. ISSN 0198-9715. Citation on page 72.

WESTERHOLT, R. **The Analysis of Spatially Superimposed and Heterogeneous Random Variables monitoramento de enchentes**. Phd Thesis (PhD Thesis) — Universität Heidelberg, 69117 Heidelberg, Germany, 3 2018. Citation on page 110.

WESTERHOLT, R.; RESCH, B.; ZIPF, A. A local scale-sensitive indicator of spatial autocorrelation for assessing high- and low-value clusters in multiscale datasets. **International Journal of Geographical Information Science**, Taylor & Francis, v. 29, n. 5, p. 868–887, 2015. Citation on page 50.

WESTERHOLT, R.; STEIGER, E.; RESCH, B.; ZIPF, A. Abundant topological outliers in social media data and their effect on spatial analysis. **PLOS ONE**, Public Library of Science, v. 11, n. 9, p. 1–31, 09 2016. Citation on page 72.

XIAO, Y.; HUANG, Q.; WU, K. Understanding social media data for disaster management. **Natural Hazards**, v. 79, n. 3, p. 1663–1679, Dec 2015. ISSN 1573-0840. Citations on pages 42, 58, 71, 73, 83, and 96.

XU, D.-L. An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis. **Annals of Operations Research**, v. 195, n. 1, p. 163–187, May 2012. ISSN 1572-9338. Citations on pages 38 and 50.

YANG, X.; XIE, X.; LIU, D. L.; JI, F.; WANG, L. Spatial interpolation of daily rainfall data for local climate impact assessment over greater sydney region. **Advances in Meteorology**, v. 2015, p. 1–12, 2015. ISSN 1687-9309. Citation on page 60.

ZHOU, X.; CHEN, L. Event detection over twitter social media streams. **The VLDB Journal**, v. 23, n. 3, p. 381–400, Jun 2014. ISSN 0949-877X. Citations on pages 28 and 71.

THEORETICAL BACKGROUND

A.1 Geographic information science

Geographic Information Science (GIScience) is a field of study that examines the nature of geographic information and the development and use of theories, methods, technologies, and data for understanding geographic processes, relationships, and patterns on different geographical scales ([MARK, 2003](#); [GOODCHILD, 2010](#); [GOODCHILD, 1992](#)). GIScience lies behind the Geographic Information Systems (GIS) and involves much more than the technology, encompassing a wide array of disciplines, such as geography, cartography, geodesy, statistics, and computer science. It still can be seen on a social frame that encompasses social disciplines (e.g., spatial cognition and social science), addressing issues of reasoning and meaning of geographic objects, people, and places, and their impacts on society. Therefore, GIScience is an inter- and transdisciplinary research area supported by a technical and social (non-technical) body of knowledge.

A.2 Spatial data analysis

Seen from the perspective of geographic processes, relationships, and patterns, spatial data analysis is a central topic in GIScience and deals with “a collection of techniques and models that explicitly use the spatial referencing associated with each data value or object that is specified within the system under study” ([HAINING, 2003](#)). Such techniques and models are essential for making and evaluating assumptions about data and their spatial relationships.

Several spatial data analysis workflows form the core of GIScience; nonetheless, they usually share two fundamental concepts: spatial autocorrelation and spatial heterogeneity ([ANSELIN, 1988](#)). While the former refers to the degree of dependence to which one object is similar to other nearby objects, the latter refers to the spatial structure of such objects. Questions such as “how much does the economic of one neighborhood influence another?” and

“Are economically poor neighborhoods concentrated or scattered within the city?” could be investigated through spatial autocorrelation and spatial heterogeneity, respectively. Both concepts are important in any spatial data analysis, including mining of thematic spatio-temporal patterns from social media data.

Spatial autocorrelation

In the most general sense, data are obtained for observations that are ordered in space or in space and time (ANSELIN, 1988, p. 11). For example, census data are collected for census tracts, usually every ten years. Sometimes, the observations are aggregated into arbitrary spatial units of analysis. For example, rainfall measurements are obtained from a set of rainfall gauges and later such measurements are aggregated into spatial units, such as neighborhoods and cities. Regardless of the level of aggregation, such data or spatial units are more or less similar as a function of the distance between them. Such a assumption is based on the first law of geography, which states that “everything is related to everything else, but near things are more related than distant things” (TOBLER, 1970, p. 236).

In light of this, spatial autocorrelation can be considered a functional relationship between what happens at one point or area in space and what happens elsewhere. In other words, adjacent or close points/areas may be subject to (dis)similar contextual influences, such as socioeconomic and meteorological factors. Several spatial autocorrelation measures exist (GETIS, 2007), such as Moran’s I and I_i , Geary’s c , and Getis and Ord’s G , G_i and G_i^* . Each measure evaluates the functional relationship in different application scenarios. These include the assessment of influences of distance effects, of the roles of geometry and topology, or of the impact that individual geographic features have on spatial processes (WESTERHOLT, 2018; GETIS, 2007). Moreover, such measures can be categorized into two groups, global spatial autocorrelation measures (e.g., Moran’s I , Geary’s c , and Getis and Ord’s G) and local spatial autocorrelation measures (e.g., Moran’s I_i and Getis and Ord’s G_i and G_i^*). In this doctoral research, global and local Moran’s were used as measures of spatial autocorrelation (see Chapter 2).

Moran’s I

Moran’s I is the most common spatial autocorrelation measure, which explains the overall distribution of a specific process and whether the process presents clustering characteristics in a given space. It is widely used because it has some advantages over other spatial correlation measures, such as higher statistical power, less affected by data outliers, and more robust against configurational outliers in the spatial arrangement (how space is divided into spatial units).

The Moran's I measure is given as:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{A.1})$$

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$$

where x_i and x_j are observations on the variable of interest at locations i and j , $\bar{x}$ is its mean, n is the total number of observations, and w_{ij} are the elements of the spatial weights matrix, which embodies a prior notion of the neighbor structure of the observations. The term S_0 corresponds to the sum of all the spatial weights.

For ease of interpretation, the weights w_{ij} in Equation A.1 may be in binary neighbourhood matrix form. However, other spatial configurations and weights can be used (ANSELIN, 1988, pp. 16–21).

$$w_{ij} = \begin{cases} 1, & \text{if } i, j \text{ are adjacent neighbors (i.e., if they have a common border)} \\ 0, & \text{otherwise} \end{cases}$$

Moran's I computes the degree of spatial autocorrelation between neighboring values of a variable of interest. It varies from -1 (perfect negative autocorrelation) to 1 (perfect positive autocorrelation), where 0 represents perfect spatial randomness (no spatial autocorrelation). Figure 30 illustrates the degrees of spatial autocorrelation.

Figure 30 – Illustrations of the degree of spatial autocorrelation.

Source: Elaborated by the author.

Nonetheless, Moran's I can only explain the clustering effects in a given space without clearly highlighting the clustering regions. For accurately identify the clustered regions, its local version should be used. By analogy, the local version of Moran's I (Moran's I_i) allows the decomposition of global Moran's I into its local constituents (ANSELIN, 1995):

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2 / (n-1)} \sum_{j=1}^n w_{ij} (x_j - \bar{x}) \quad (\text{A.2})$$

Spatial heterogeneity

Spatial heterogeneity is related to the lack of instability of the variable of interest over space, i.e., values of the variable of interest vary abruptly among nearby places, or similar values of the variable of interest are in areas far from each other. Figure 31 illustrates two forms of spatial heterogeneity presented by Fahrig *et al.* (2011): compositional heterogeneity and configurational heterogeneity. While compositional heterogeneity refers to the abrupt change in the values of nearby areas (Figure 31 (y-axis)), configurational heterogeneity refers to a more complex spatial arrangement (Figure 31 (x-axis)).

This uneven spatial distribution is often due to location-specific effects, such as topography and socioeconomic factors. Moreover, It can manifest in the form of different response functions or systematically varying parameters, such as heteroskedasticity, random coefficient variation, switching regressions, and spatial arrangement configuration (ANSELIN, 1988). Regardless of how spatial heterogeneity manifests, it shows that a single global relationship for an overall study region may not adequately reflect outcomes in any given location of the study region.

Figure 31 – Illustration of the two major axes of spatial heterogeneity: compositional and configurational heterogeneity. Each large square is an area and different colors represent different ranges of values (or quantiles). Compositional heterogeneity increases as values/areas belong to different intervals of values (or quantiles). Configurational heterogeneity increases with the increasing complexity of the spatial arrangement.

Source: [Fahrig *et al.* \(2011\)](#).