Estimating Vulnerability to Extreme Events in Urban Areas: From Microdata to City-Wide Indicators

Eduardo A. Reolon[1], **Ricardo Lüders**[1], **Luiz Gomes-Jr.**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80230-901 – Curitiba – PR – Brazil

eduardoreolon@alunos.utfpr.edu.br, {luders,lcjunior}@utfpr.edu.br

Abstract. *During extreme event crises, such as floods, it is crucial to have accurate information about the most vulnerable populations. This study proposes a cost-effective methodology for estimating flood vulnerability by integrating microdata from household-level social records with macrodata derived from demographic and socioeconomic indicators. This approach allows for constructing a regional vulnerability index, which enables identifying high-risk areas even without complete microdata coverage. The method overcomes the challenge of limited data availability by leveraging existing public data sources and a regression model. The results of a case study in Curitiba, Brazil, show the potential to support the prioritization of regions for emergency response and to guide the development of public policies and mitigation strategies based on social and demographic criteria. This replicable methodology can be adapted to other cities facing budget constraints as a practical tool for mapping vulnerabilities and enhancing risk management.*

1. Introduction

Climate change has made flooding more frequent and intense, disproportionately affecting the most vulnerable populations. A study by the Porto Alegre Hub of the INCT Observatory of Metropolises revealed that, during floods in Rio Grande do Sul, the poorest areas with a higher concentration of black populations were the hardest hit, highlighting a pattern of inequality that repeats across various regions of Brazil [Augustin 2024]. Therefore, it is crucial that intervention and recovery programs take into account the socioeconomic specificities of the affected populations, promoting an equitable and sustainable recovery.

Flooding brings a series of significant damages, directly affecting the lives of thousands of people around the world. In addition to the tragic loss of human life, these catastrophic events destroy essential infrastructure, including homes, schools, and bridges. Floods also leave many people exposed to waterborne diseases. The situation is exacerbated in developing countries, where rapid urbanization and illegal construction in high-risk areas increase the vulnerability of local populations, making damage management and mitigation an even more challenging task [Rehman et al. 2021].

Traditional approaches to disaster risk management increasingly emphasize population vulnerability, in line with the Sendai Framework [de Brito et al. 2017], which advocates for strategies that address all dimensions of risk, including individuals' capacity to respond. Several studies [Mendonça and Buffon 2016, Debortoli et al. 2017, Rasch 2016] highlight how socioeconomic conditions, lack of infrastructure, housing quality, and access to services are key factors in assessing vulnerability and anticipating future risk scenarios. Recognizing and addressing these aspects is essential to reducing inequalities and mitigating the impacts of floods.

The present work proposes a method for estimating flood vulnerability by integrating microdata, from household-level social records, with macrodata, derived from demographic and socioeconomic indicators. While microdata offer detailed insights into living conditions, they are often limited in coverage. By combining them with broader indicators, this approach enables a more comprehensive understanding of vulnerable regions and reduces the costs associated with extensive data collection across the entire city. This methodology complements the ICARUS system [Fernandez and Splendore 2021], which focuses on flood detection and communication network management during emergencies, by incorporating social vulnerability into the analysis.

The proposed methodology combines a machine learning model, such as XGBoost, with socioeconomic and demographic data to estimate flood vulnerability at a regional level. The approach allows for identifying more vulnerable populations by utilizing indicators such as income, property ownership, and family composition. The results from the case study in Curitiba demonstrate the method's potential to aid in resource allocation and in the planning of emergency response and mitigation actions.

This paper is organized as follows: In Section 2, essential concepts are discussed, such as the different types of floods and main data sources. Section 3 presents the criteria adopted to assess vulnerability based on social and demographic data. Section 4 presents a case study applying the method to Curitiba and analyzing the resulting vulnerability estimates. Section 5 summarizes the results achieved, highlighting the system's contribution to mitigating flood vulnerabilities.

2. Background and Related Work

2.1. Flood and related definitions

The concepts of flood, inundation, and waterlogging represent distinct hydrological phenomena resulting from excessive precipitation. The term *bankfull* denotes the condition in which a drainage channel reaches its maximum capacity without overtopping its banks. When this threshold is surpassed, water spreads into adjacent low-lying areas, characterizing an inundation. In urban environments, inadequate drainage infrastructure may lead to the accumulation of surface water, a process commonly referred to as urban flooding or waterlogging [Rodrigues 2023]. More broadly, flooding is defined as a condition in which land is submerged by water, especially as a result of intense rainfall, encompassing various scenarios across both rural and urban settings.

Our methodology focuses on assessing the socioeconomic and demographic factors that influence vulnerability to these phenomena in urban environments, as they directly impact a population's capacity to cope with and recover from floods.

2.2. Data proxies for underrepresented areas

Obtaining data on vulnerable populations is a significant challenge, directly impacting planning and response actions in flood situations. The Relative Wealth Index (RWI), proposed by Guanghua Chi et al. (2022) [Chi et al. 2022], is a measure used to estimate wealth distribution across different regions, providing granular insights into socioeconomic disparities. Chi et al. (2022) conducted an analysis using a neural network, incorporating features such as average road density, average elevation, average annual precipitation, number of mobile cellular towers, WiFi access points, mobile devices of various types, and principal components extracted from satellite imagery. These data were essential for training the model and identifying socioeconomic patterns.

The "Population Density Explorer" [Facebook 2024] provided by Facebook's Data for Good initiative is an interactive tool designed to help users explore high-resolution population density maps. This tool allows individuals to access and download summary statistics for various geographic regions, including custom boundaries. The data, which comes from Facebook's vast user base and its partnerships, is especially useful for humanitarian organizations, governments, and researchers in responding to challenges like natural disasters and social crises. The maps assist in decision-making for resource allocation, such as determining the locations for critical services, and have been applied in diverse contexts like COVID-19 response and climate change mitigation. Additionally, the tool presents demographic breakdowns, including the total number of men, women, children under 5 years old, youth aged 15-24, and elderly individuals aged 60 and above.

In this work, we leverage these types of proxy data, such as the Relative Wealth Index and high-resolution population maps, to address data scarcity and improve the breadth of our vulnerability assessments.

2.3. Vulnerability Score

Vulnerability parameters associated with the sociodemographic characteristics of families enables an understanding of how socioeconomic conditions influence the resilience of populations most affected by natural disasters. Density stands out as highly relevant for the calculation of vulnerability [Scalenghe and Marsan 2009], not only from the probability perspective, as a larger group of individuals increases the chance of having more people in conditions of socioeconomic fragility. The degree of soil sealing, as a consequence of population density, is directly related to the occupation pattern and the intensity with which urban areas are used [Scalenghe and Marsan 2009]. This impermeabilization will increase the amount and speed of water on the surface, significantly raising the risk of flooding [Scalenghe and Marsan 2009].

Some groups will require greater attention and special care during critical moments. Children, generally of shorter stature, are more susceptible to drowning [Rasch 2016], while elderly individuals over 65 years old and people with disabilities typically have less knowledge to deal with adverse situations or will simply depend on specific rescue conditions [Debortoli et al. 2017].

While of little relevance at the moment of need, when efforts are not immediately rewarded financially, income presents itself as a crucial factor in preparation and post-catastrophe recovery [Rasch 2016]. Considered by economists as a form of capital, Property Ownership is part of the family's assets, enhancing its resilience during times of need [Rasch 2016]. The Brazilian Institute of Economics (FGV IBRE) [Ibre 2024], in the year of this study, considers poverty as a per capita income below R$667, and extreme poverty as a per capita income below R$209.

Based on these established criteria, our methodology integrates socioeconomic and demographic indicators like income, age, and property ownership to create a comprehensive vulnerability score for families and regions.

2.4. Related Work

Table 1 summarizes the related work on vulnerability assessment methodologies. In the Brazilian context, Mendonça et al. [Mendonça and Buffon 2016] explored the concept of spatial resilience in the Cajuru neighborhood of Curitiba-PR, revealing that disaster-prone areas often coincide with socially vulnerable zones. Their analysis considered seven equally weighted categories: literacy, income, street garbage, irregular occupation, sewage or septic tank system, garbage collection, and water supply. On a broader scale, Debortoli et al. [Debortoli et al. 2017] examined Brazil's vulnerability to floods and landslides in the context of climate change, using indicators such as the MHDI and GINI index to highlight the northern and northeastern regions as the most vulnerable. Similarly, Rasch [Rasch 2016] developed a flood vulnerability index for 1,276 Brazilian municipalities, integrating demographic, socioeconomic, and housing indicators to pinpoint high-risk areas. Building on these contributions, our work adopts key indicators such as property ownership, per capita income, construction material, and family composition. This approach focuses on a methodology tailored for regions with scarce data, combining incomplete and proxy information to produce a broader assessment of flood vulnerability.

Table 1. Comparison of Vulnerability Assessment Methodologies

Aspect	Mendonça et al. [Mendonça and Buffon 2016]	Debortoli et al. [Debortoli et al. 2017]	Rasch [Rasch 2016]	Our Proposed Method
Geographic Scope	Local (Cajuru neighborhood, Curitiba)	National (Brazilian municipalities)	National (1,276 Brazilian municipalities)	Local (City of Curitiba)
Data Sources	Socioeconomic data, Seven equally weighted categories	Municipal Human Development Index (MHDI), six indicators and GINI index	Wide range of demographic, socioeconomic, and housing indicators	Microdata (household-level social records) and Macrodata (demographic and socioeconomic indicators)
Main Goal	Define spatial resilience and apply it to a specific neighborhood	Assess vulnerability to floods and landslides at a national scale	Present a flood vulnerability index for Brazilian municipalities	Estimate flood vulnerability by integrating microdata and macrodata to overcome incomplete information
Methodology	Social vulnerability mapping using cartographic synthesis with map algebra and multicriteria analysis	Calculation and normalization of socioeconomic indicators (0–1 scale) for integration into multidimensional vulnerability analysis	Incorporates various indicators to create a flood vulnerability index	Combines individual family scores (V_f) with a regression model (XGBoost) to extrapolate vulnerability
Key Contribution	Highlights the link between vulnerable areas and high-risk zones	Projects increased vulnerability due to climate change	Identifies high-risk municipalities and variables contributing to vulnerability	Proposes a cost-effective, replicable methodology for data-scarce regions

3. Methodology

Given the high costs and logistical difficulties of collecting new data for city-wide urban vulnerability assessments, especially when available information is incomplete, we propose a practical and low-cost methodology that draws on existing data from municipal departments as well as broader demographic and socioeconomic indicators. In the Brazilian context, Francisco Mendonça et al. [Mendonça and Buffon 2016] define spatial resilience as “the ability of an environment or society to return to its previous state after any impact.” Families with lower resilience, due to socioeconomic limitations and insufficient infrastructure, are more exposed to the impacts of floods. Building on this premise, our approach uses geographically limited microdata to extrapolate a comprehensive vulnerability model for the entire urban area. This approach is considered low-cost because it leverages existing data and avoids the high expense and complexity of large-scale, new data collection.

Furthermore, available microdata might not truly represent all the diversity of the population. For example, data could overemphasize low-income populations while ignoring middle or high-income areas. To fix this, a city might need to manually assign vulnerability levels to certain regions based on expert judgment, qualitative assessments, or other supporting information. This helps reduce bias and ensures our extrapolated vulnerability index accurately reflects the city’s diverse social and spatial contexts.

3.1. Proposed Method

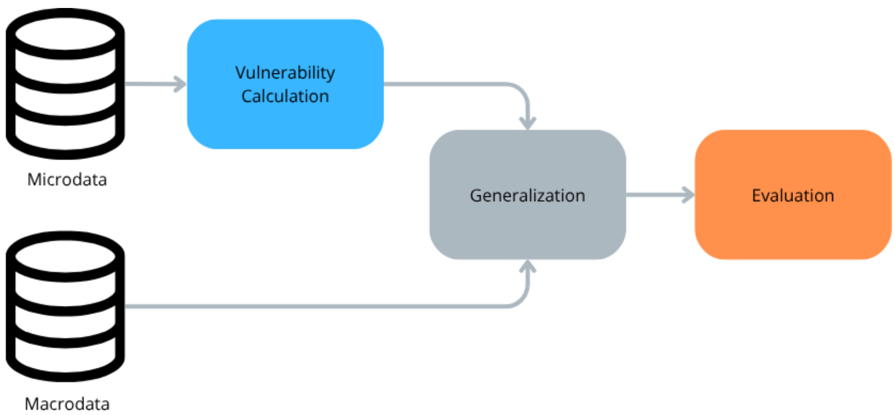

Figure 1. Methodology Overview

Figure 1 outlines the proposed urban vulnerability assessment methodology. It illustrates the processing of Microdata for vulnerability calculation, its integration with Macrodata, subsequent generalization, and an evaluation phase. These elements are described in the following sections.

3.2. Aggregated Family Vulnerability Using Microdata

The vulnerability assessment is structured in two main steps that combine microdata and spatial aggregation. First, a Family Vulnerability (FV) score is calculated individually for each household, using a composite index based on socioeconomic indicators. This score reflects the specific conditions of each family in relation to flooding vulnerability. Second, these individual scores are aggregated by region to obtain a Regional Vulnerability ($\text{FV}_{\text{region}}$) measure, enabling spatial analysis and comparison across different areas.

1. **Calculation of Family Vulnerability (V_f):**
   The vulnerability of each family is individually quantified using a composite score

ranging from 0 to 1 (after normalization). This score is a weighted average of multiple indicators given by (1):

$$V_f = \frac{TI + RC + MC + 2 \cdot MS}{5} \tag{1}$$

where:

- **Type of Property (TI):**
  - 0 for owned properties
  - 1 for rented properties
- **Per Capita Income (RC):** Values assigned according to income per person:
  - 1 if income $\leq$ 209 BRL
  - 0.8 if income $\leq$ 667 BRL
  - Linearly decreasing to 0 for income up to 2 minimum wages
- **Construction Material (MC):**
  - 0 for masonry (brick/concrete)
  - 0.5 for wooden constructions
  - 1 for other, lower quality materials
- **Family Composition (MS):** Combines the age dependency ratio and presence of vulnerable groups:
  - **Age Dependency Score:** ratio of dependents (under 15 or over 64) to other family members
  - **Sensitive Groups Score:**
    - +0.3 if at least one person has a disability
    - +0.3 if at least one child is present
    - +0.3 if at least one elderly person is present
    - +0.3 if the household is headed by a woman with at least one child and no partner
  - The total MS value is capped at a maximum of 1

2. **Regional Vulnerability V_r:**
The regional vulnerability score (2) is calculated as the average of N individual family scores within the region:

$$V_r = \frac{1}{N_f} \sum_{i=1}^{N_f} V_{f,i} \tag{2}$$

where:

- V_r denotes the regional vulnerability score;
- N_f is the total number of families within the region;
- $V_{f,i}$ represents the family vulnerability score for the i-th family.

3.3. Macrodata Regression for Regional Vulnerability

This methodological step, Generalization, aims to build a predictive model. The model learns the relationship between the aggregated socioeconomic indicators (macrodata) and the vulnerability index previously calculated from the microdata. The purpose is to create a function capable of estimating vulnerability for locations where only macrodata is known. This task is formalized as a regression problem.

The general relationship can be expressed by the following formula:

$$V_r = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_{ir} \tag{3}$$

Where:

- V_r denotes the estimated vulnerability index for region r;
- X_{ir} represents the value of socioeconomic and demographic variables (e.g., RWI, percentage of women, elderly population) for region r;
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients to be estimated based on observed data.

While the general formula describes a linear relationship, this estimation task can be performed using various machine learning models. Techniques such as Random Forest and XGBoost, for example, can capture complex, nonlinear patterns among the macrodata indicators, potentially leading to more accurate and robust predictions of regional vulnerability.

To assess the effectiveness of the chosen model, its predictions of regional vulnerability can be evaluated using appropriate performance metrics for regression tasks, such as R-squared (R^2) and Mean Squared Error (MSE). Additionally, the resulting spatial patterns of vulnerability can be compared against external references, like expert opinions or alternative datasets even if at a lower resolution, to ensure the model captures meaningful and consistent insights.

4. Case Study: Application of the Method in the City of Curitiba

4.1. Geographical Data

In a previous work related to this proposal, Heron Gomes Fernandez et al. [Fernandez and Splendore 2021] discretized the city of Curitiba into 600-square-meter hexagons. In order to facilitate processing and enable future integration with that work, the same areas were chosen. This approach also ensured that no personal address data of the families was used, but only the area to which they belong, guaranteeing privacy and data protection.

4.2. Socioeconomic Microdata

Socioeconomic and demographic data from the Housing Company of Paraná (COHAPAR) [COHAPAR 2024] were used, collected between 2016 and 2024, covering 22,346 families and 31,515 people. COHAPAR authorized the use of this data, recognizing the benefit that its analysis can bring to society, ensuring that no personal information was used, as the data was properly anonymized.

The analysis used a selection of variables relevant to assessing social vulnerability, considering housing conditions, family income, and demographic characteristics. Table 2 summarizes the fields used in the proposed method.

4.3. Socioeconomic Macrodata

The Relative Wealth Index (RWI) is calculated based on a 2.4 km grid cell resolution, resulting in 93 mapped points within the city of Curitiba. Figure 2 shows the results of a Kriging interpolation process, with scores ranging from below 0.5 to above 0.8.

Families	Members
imovel_proprio (boolean)	idade (integer)
renda_total (float)	deficiencia_alguma (boolean)
numero_membros (integer)	sexo (char)
tipo_construcao (enum)	parentesco (string)

Table 2. Selected Fields Used in the Vulnerability Analysis

Figure 3 shows the population density of Curitiba's regions using data from Facebook's *Population Density Explorer*. The dataset also includes information on the distribution of children and the elderly, enabling a more detailed analysis of vulnerable groups and supporting targeted public policy actions.

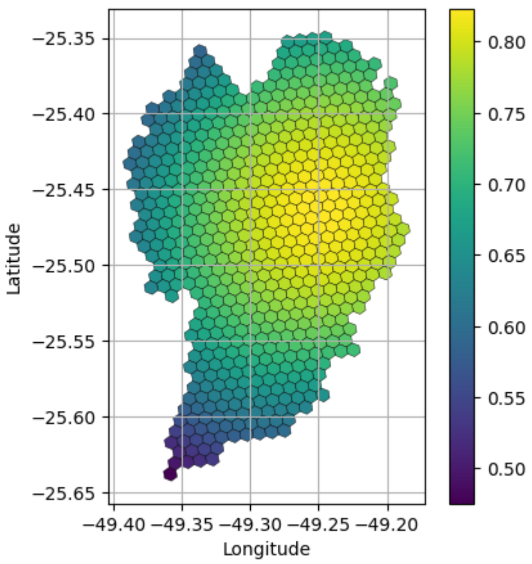

Figure 2. RWI index discretized in micro areas

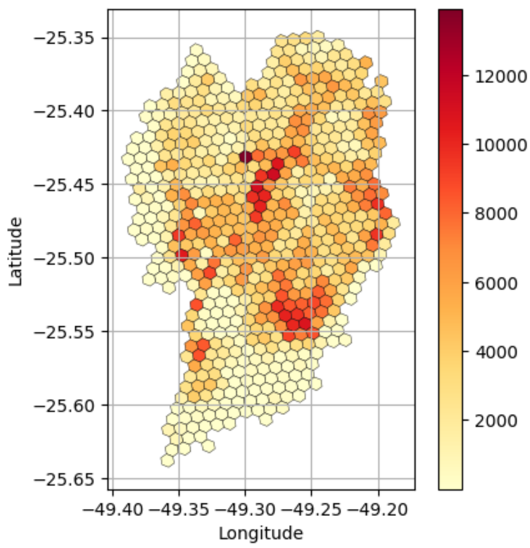

Figure 3. Total population by facebook

4.4. Regression Model

To ensure good coverage of the microdata, a selection of 50 areas was made based on a range metric calculated as the ratio between the number of families in the COHAPAR [COHAPAR 2024] registry and the estimated population of each area, according to Meta's high-resolution population dataset [Facebook 2024]. Within these 50 areas, some received manually assigned vulnerability scores based on an IBGE map that classifies urban areas in Curitiba from A (most developed) to K (least developed) [IBGE 2021]. Fifteen areas labeled as A or B on this scale were assigned a fixed vulnerability score of 0.05. These areas were considered to be sufficiently represented and were included in the selection to compensate for the lack of data on low-vulnerability regions in the COHAPAR dataset.

The XGBoost model presented strong predictive capabilities in estimating the vulnerability scores, utilizing the Richness Wealth Index (RWI), percentage of women, and percentage of elderly individuals (60+) as input variables. After refining this set of input variables, the model achieved an R^2 of 0.722 and a Mean Squared Error (MSE) of 0.0028.

Figure 4 shows the vulnerability based on household-level microdata, while Figure 5 displays the predicted vulnerability scores generated by the XGBoost model. Both maps use the hexagonal grid, where lighter colors indicate lower vulnerability and darker colors represent higher vulnerability. Figure 6 shows the IBGE classification of urban typologies, in which colors from blue to red represent a gradient from better to worse living conditions. Despite differences in data sources and classification criteria, the spatial patterns observed across the three maps are similar, suggesting that the proposed method captures relevant aspects of the city's vulnerability profile with greater detail, which can better guide decision-making processes.

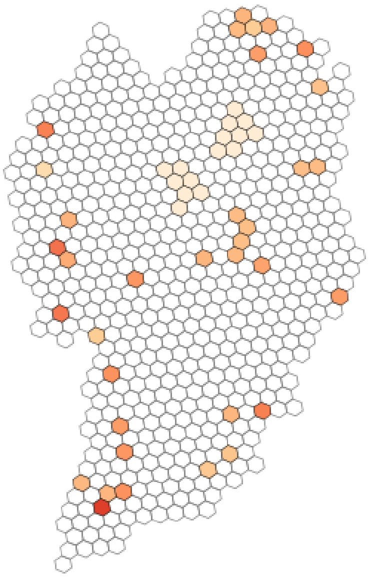

Figure 4. Vulnerability based on microdata

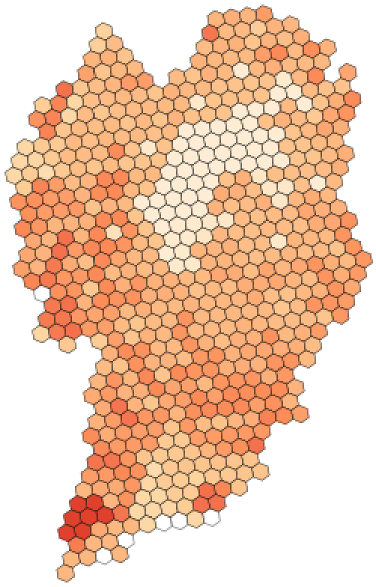

Figure 5. Predicted vulnerability scores across Curitiba

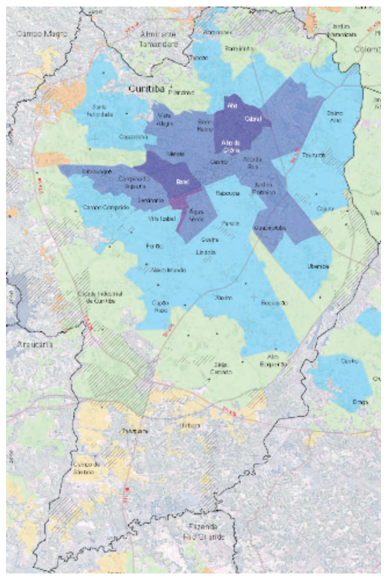

Figure 6. IBGE urban classification reference

5. Conclusion

The methodology developed in this study proved effective in identifying areas and populations vulnerable to flood impacts by combining data from assisted families with publicly available information. Constructing a vulnerability index allowed for the extrapolation of local characteristics to broader regions, demonstrating potential to support decision-making in urban planning, resource allocation, and emergency response. By focusing on accessible data sources and replicable techniques, the proposed approach offers a practical, low-cost solution that can be easily adapted by other municipalities facing budgetary and data-related challenges.

Despite its potential, the study has some limitations that point to directions for future research. The main limitation is sampling bias, as the analysis relied on data from families enrolled in social programs, which may limit the generalization of the results to wider populations. Additionally, the current model does not consider geographic information on areas' susceptibility to flooding, only focusing on social aspects of vulnerability. Future research could expand the dataset to include a more diverse sample and integrate geographic risk data, enabling a more comprehensive analysis and a more robust model for disaster risk management.

Acknowledgments

Projeto SocialNet/Fapesp (2023/00148-0), and CNPq (444724/2024-9, 441444/2023-7).

References

Augustin, A. (2024). Núcleo Porto Alegre analisa os impactos das enchentes na população pobre e negra do Rio Grande do Sul. *Observatório das Metrópoles*.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences (PNAS)*, 119(3):e2113658119.

COHAPAR (2024). https://www.cohapar.pr.gov.br/.

de Brito, M. M., Evers, M., and Höllermann, B. (2017). Prioritization of flood vulnerability, coping capacity and exposure indicators through the delphi technique: A case study in Taquari-Antas basin, Brazil. *International Journal of Disaster Risk Reduction*, 24:119–128.

Debortoli, N. S., Camarinha, P. I. M., Marengo, J. A., and Rodrigues, R. R. (2017). An index of Brazil's vulnerability to expected increases in natural flash flooding and landslide disasters in the context of climate change. *Natural Hazards*, 86:557–582.

Facebook (2024). Data for good: Visualizations. https://dataforgood.facebook.com/dfg/visualizations.

Fernandez, H. G. and Splendore, P. R. (2021). Sistema de identificação automática de riscos hidrometeorológicos com retroalimentação e reestruturação autônoma da infraestrutura de comunicação. Bachelor's thesis, UTFPR.

IBGE (2021). Tipologia intraurbana: Espaços de diferenciais socioeconômicos nas concentrações urbanas do Brasil. https://geoftp.ibge.gov.br/organizacao_do_territorio/tipologias_do_territorio/tipologia_intraurbana/Tipologia_Intraurbana.pdf.

Ibre, P. (2024). Informações econômicas. https://portalibre.fgv.br/.

Mendonça, F. and Buffon, E. A. M. (2016). Resiliência socioambiental-espacial urbana à inundações: possibilidades e limites no bairro Cajuru em Curitiba (PR). *Início / Arquivos*, 12(19).

Rasch, R. J. (2016). Assessing urban vulnerability to flood hazard in Brazilian municipalities. *Environment and Urbanization*, 28(1):1–16.

Rehman, A., Akhtar, N., and Alhazmi, O. H. (2021). Formal modeling, proving, and model checking of a flood warning, monitoring, and rescue system-of-systems. *Scientific Programming*, 2021.

Rodrigues, G. T. (2023). Modelo de previsão de inundações urbanas: uma abordagem baseada em dados da bacia hidrográfica Ribeirão dos Padilhas, Curitiba (PR). Master's thesis, PUC-PR.

Scalenghe, R. and Marsan, F. A. (2009). The anthropogenic sealing of soils in urban areas. *Landscape and Urban Planning*, 90(1-2):1–10.