Estimating intraurban socioeconomic status using users'interactions registered on digital data

Eduardo Segundo Cruz Ramirez

Facultad de Ingeniería en Electricidad y Computación - FIEC
Escuela Superior Politécnica del Litoral - ESPOL

This dissertation is submitted for the degree of
Doctor of Applied Computer Science

Guayaquil, 2024

PhD. Candidate	**Eduardo Cruz Ramírez** Escuela Superior Politécnica del Litoral
Director	**PhD. Carmen Vaca Ruiz** Escuela Superior Politécnica del Litoral
Co-Director	**PhD. Monica Villavicencio Cabezas** Escuela Superior Politécnica del Litoral
Thesis Committee	**PhD. Daniel Ochoa Donoso** Escuela Superior Politécnica del Litoral **PhD. Luis Eduardo Mendoza** Escuela Superior Politécnica del Litoral **PhD. Francisco Moreno Arboleda** Universidad Nacional de Colombia **PhD. Hugo Alatrista Salas** Léonard de Vinci, Pôle Universitaire Research Center

Declaración Expresa

Yo Eduardo Segundo Cruz Ramírez acuerdo y reconozco que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL. La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi innovación, de ser el caso. En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Eduardo Segundo Cruz Ramirez

Guayaquil, 2024

Evaluadores

PhD. Carmen Vaca Ruiz
Escuela Superior Politécnica del Litoral

PhD. Monica Villavicencio Cabezas
Escuela Superior Politécnica del Litoral

Acknowledgements

Thank God for having excellent academic and personal guides who have allowed me to complete this work, such as my director, Carmen Vaca, and my co-director, Mónica Villavicencio. Their knowledge and support were essential to achieve this work. This work is dedicated to my parents who, with their effort and dedication, have guided me to achieve the goals that I have set for myself in my personal and professional development. Finally, I want to thank my family, friends, and colleagues, who were always watching out for me to achieve this work.

Abstract

The thesis addresses the challenge of estimating socioeconomic status (SES) at an intraurban level using digital data sources. Traditional methods for measuring SES, such as censuses and surveys, are often limited by their infrequency and coarse spatial granularity, which hinders timely and accurate assessments, especially at the neighborhood level. The study proposes leveraging alternative digital data sources, including mobile phone top-up transactions and supermarket purchase data, to model and predict SES, providing the potential for more frequent, cost-effective, and spatially granular analysis. The research focuses on urban neighborhoods in Ecuador, aiming to develop machine learning models that can accurately predict Neighborhood SES (NSES).

The research employs two machine learning models: a Regression Model using mobile phone top-up transactions and a Graph Neural Network (GNN) Model using supermarket transaction data. The first model focuses on linear relationships between variables derived from top-up transaction data and NSES. The model is designed to estimate the NSES by aggregating the average denomination and the denomination diversity at the neighborhood level. The second model leverages the complex, non-linear relationships inherent in supermarket transactions. The GNN model transforms these transactions into a graph representation, where items purchased together are linked, and the frequency and diversity of these links are analyzed to infer SES. The model is particularly suited for capturing the socioeconomic patterns that emerge from the co-purchase behaviors of individuals within a neighborhood.

Both models demonstrate significant predictive power in estimating SES at the intraurban level. The Regression Model achieves a prediction accuracy of up to 74%. This model is particularly effective in identifying the relationship between average top-up denomination and neighborhood SES, with higher denominations indicating wealthier neighborhoods. The GNN Model outperforms the Regression Model, achieving a prediction accuracy of up to 91%. The GNN model is able to model the intricate patterns of co-purchases within neighborhoods, allowing for a more detailed and accurate representation of NSES. The results highlight the potential of digital data sources as viable alternatives to complement traditional SES measurement methods.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

In 2015, the United Nations (UN) members adopted the Agenda for Sustainable Development as a universal call to action to end poverty, protect the planet, and ensure that by 2030, all people enjoy a better and more sustainable planet [9]. The agenda comprises 17 Sustainable Development Goals (SDGs) [1], which call for urgent action by all countries. SDG 1 is to end poverty in all its forms everywhere, which remains the most pressing problem the world faces. Despite efforts to tackle poverty, studies suggest that the number of people who live in extreme poverty and struggle to access basic needs such as health, education, water, and sanitation will remain above 600 million in 2030 [10]. Therefore, it is essential for governments to assess regularly at various levels (national, urban, and intra-urban) the socioeconomic status (SES), which is an economic and social combined measure of a person or group's economic and social position in relation to others based on income, occupation, housing conditions, wealth levels, and education to make informed decisions on how to allocate resources according to the needs of different geographical areas.

The emergence of the COVID-19 pandemic has posed a global health crisis since late 2019. The World Bank reported that the number of people living in extreme poverty has declined for the last three decades. However, the trend was interrupted in 2020, when the extreme poverty rate increased due to a worldwide economic contraction caused by the COVID-19 crisis [11]. The global extreme poverty rate increased sharply from 2019 to 2020, from 8.3 percent to 9.2 percent, which means that the number of people living on less than $1.90 a day increased from 645 to 733 million [1]. Figure 1.1 illustrates the non-pandemic, COVID-19 baseline, and COVID-19 downside projection of global extreme

[1] https://sdgs.un.org/es/goals

poverty. Moreover, the Economic Commission for Latin America and the Caribbean (ECLAC) reported that Latin America and the Caribbean are the most vulnerable regions due to the prolonged health and social crisis stemming from the COVID-19 pandemic. The extreme poverty rate rose from 13.1 % of the population in 2020 to 13.8 % in 2021 [12]. Consequently, a health crisis can especially affect people living in extreme poverty, and governments need to understand the spatial distribution with a high granularity of the most vulnerable areas to respond with resource allocation and develop economic recovery programs for the most affected.

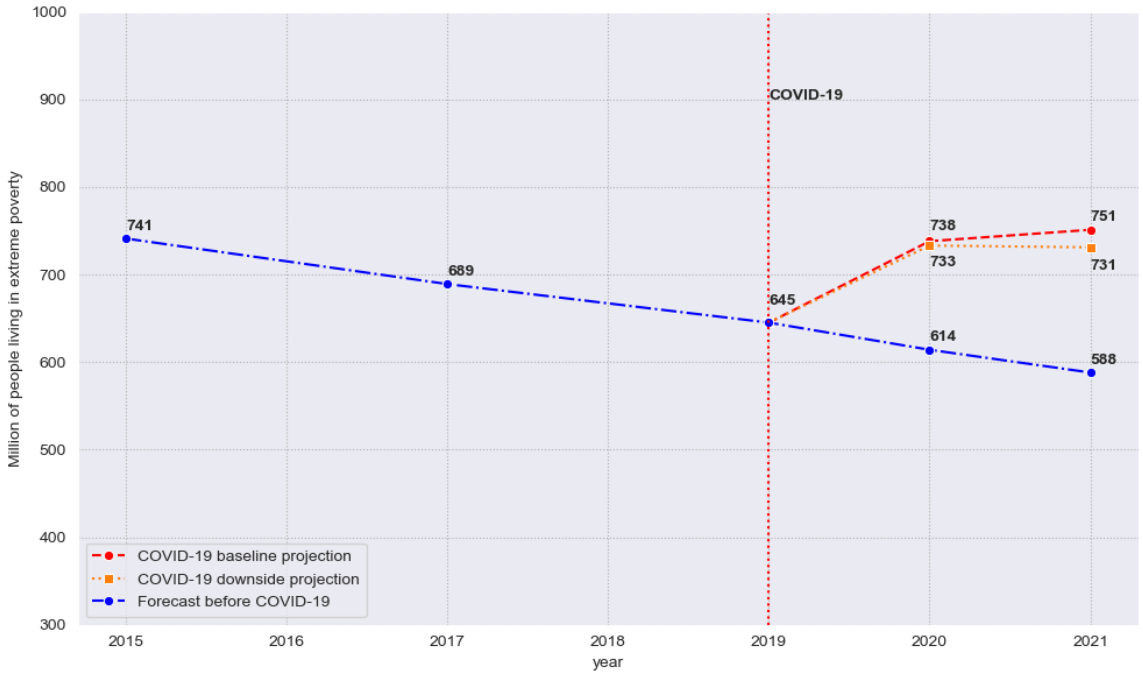


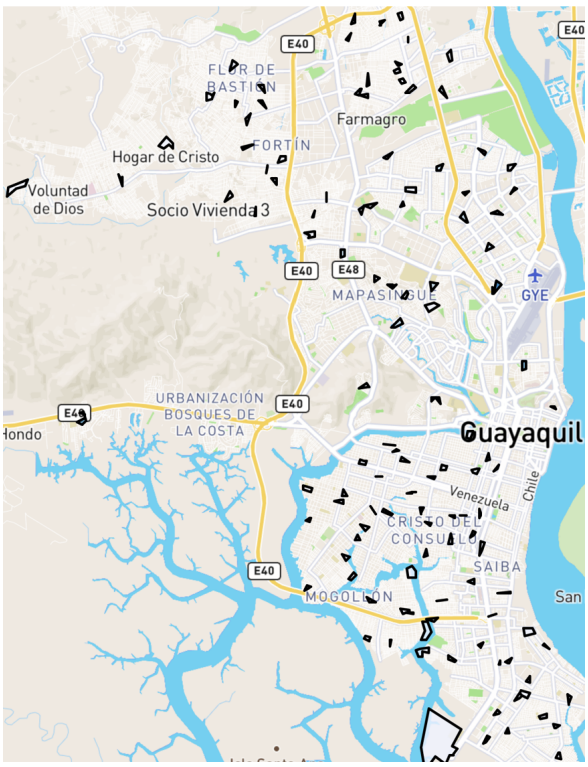
Fig. 1.1 The non-pandemic, COVID-19 baseline, and COVID-19 downside projection of global extreme poverty. [1]

Governments employ the measurement of SES as a powerful analytical tool, providing insight into individuals' quality of life and living conditions within a population. The National Statistical Institutes (NSIs) of Governments collect, measure, and update the SES. The NSIs of developing countries employ traditional data collection methods, such as censuses and surveys, to measure SES, including monetary unidimensional like Per capita income (PCI) and non-monetary multidimensional indices like the Multiple Deprivation Index (MDI). However, censuses have the disadvantages of being time-consuming and labor-intensive for NSIs. Hence, they perform these data collection methods over long-time intervals, generally ten years, causing the availability and updating of the distribution of wealth of the entire

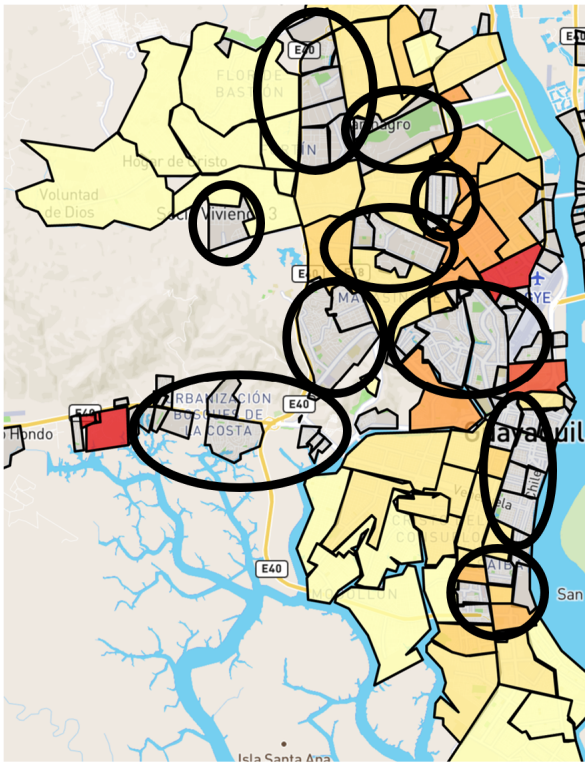
population necessary for governments to develop their economic recovery programs to be subject to these long-time intervals [13, 14, 8].

The NSIs of developing countries employ surveys that focus on a sample of the population and are performed on shorter time intervals to overcome the limitations of censuses in measuring SES. However, surveys have the disadvantage of results being regularly reported with coarse spatial granularity, generally at administrative levels 1-3 (province, canton, parish), covering multiple districts with urban neighborhoods of different socioeconomic characteristics [15–17]. For example, the NSI of Ecuador (Instituto Nacional de Estadística y Censos (INEC)) uses a spatially disaggregated sampling frame based on census tracts (around 150 dwellings) for sociodemographic research. INEC surveys the SES of a sample of census tracts to be reported quarterly at the city level. However, when governments perform an analysis of the SES at a neighborhood spatial resolution to understand the most vulnerable neighborhoods of a city to develop resource allocation programs, it can be observed that the sample of surveyed census tracts does not cover all the neighborhoods in a city for the limited population coverage. Figure 1.2 shows the surveyed census tracts and the neighborhoods for which the SES was unavailable for Guayaquil (the most populated city in Ecuador) in the first quarter of 2017. Consequently, the scarcity of reliable and updated statistics at the neighborhood level represents a significant challenge for policy-makers of governments in developing countries to design policies to allocate resources according to the needs of different geographical areas [18].

To bridge this gap, recent work has studied new alternative data sources that record digital traces of demographic characteristics, human behavior, and living conditions of the population, which, together with the use of machine learning architectures, enables new approaches to measure the SES at the intra-urban level through proxies that complement traditional methods and thus contribute to mitigating their limits. The alternative data sources include: credit/debit cards transactions [19, 20], Call Detail Records (CDRs) [2, 3], satellite imagery [4, 21, 5], social media [6], and geographic information [7, 8]. Most of these alternative data sources are readily available in wealthy nations; however, in developing countries, the access and use to these data sources are limited, or the scope is not suitable to the economic reality [22]. For example, the NSIs of developed countries employ the measure of SES extracted from digital data sources where the most remarkable economic activity of their population is recorded. The Office of National Statistics of the UK measures one real-time indicator called "Aggregate UK spending on debit and credit cards" extracted from the payments made by credit and debit card payment processors to around 100 major UK retail corporates; 59% of all payments in the UK were made using cards in 2022 and with an increasing trend [23, 24]. However, this indicator would have a different scope in



(a) Census tracts



(b) Unsurveyed neighborhoods

Fig. 1.2 (a) The surveyed census tracts and (b) the neighborhoods for which the SES was unavailable for Guayaquil in the first quarter of 2017.

developing countries where most people make cash payments. For example, the NSI of Ecuador surveyed a sample of households on preferred means of payment in 2019, and 95% of the households surveyed prefer to pay their billings in cash [25]. The exception is mobile phone data, which researchers in developed and developing countries have analyzed to understand people's daily patterns and predict SES at a small spatial resolution within a city [2, 3]. Nevertheless, other dynamics in developing countries related to mobile phone use have barely been explored, such as airtime top-up transactions, a widely used modality to purchase airtime for mobile lines. Thus, this alternative data source can be studied to understand the relationship between the population's airtime purchasing patterns and SES. In addition to airtime top-up transactions, other digital data sources where purchase transactions are recorded, such as supermarket transactions, can be explored to study the economic rhythms of the population in developing countries. Therefore, these digital data sources that record purchase transactions require a study to extract patterns that allow modeling the relationship between consumption behavior and the SES at a neighborhood level to be used as proxies that complement traditional methods and thus contribute to mitigating their limits of measuring the SES in developing countries.

1.2 Research questions

The following research questions have been formulated to tackle the limits of estimating SES at the intra-urban level:

1. Can meaningful patterns be extracted from purchase transactions recorded in digital data to assess neighborhood socioeconomic status accurately?

2. How well does users' behavior, encapsulated in digital data, reflect the spatially aggregated socioeconomic status at a neighborhood level?

1.3 Research goals

This research focuses on analyzing purchase transactions recorded in digital data to calculate metrics as a proxy to predict SES at the intra-urban level. The most important research activities are mentioned below:

1. To build a machine-learning architecture that estimates socioeconomic status at the neighborhood level using users' interactions registered on digital data.

2. To characterize urban neighborhoods of a city using estimated and aggregated socioeconomic status at the intra-urban level, allowing the establishment of rankings between the different urban neighborhoods.

Chapter 2

Literature Review

An extensive body of research has used surveys and various digital sources, such as Call Detail Records (CDRs), satellite imagery, geographic information, and social media, to extract features for innovative models that estimate SES. This work will discuss the most recent research efforts that employ digital sources to predict SES at the intra-urban level.

2.1 Call Detail Records (CRDs) to predict SES

The expansion of mobile phone use in developing countries has provided a rich and broad source of information about regions' socioeconomic characteristics. CDRs, which are records of mobile usage like calls and text messages obtained from mobile network operators, provide mobile phone activity data that can be used to model the relationship between customer behavior and SES.

Hernandez et al. [2] explored the potential of utilizing CDRs from cell phones as a novel and cost-effective method for estimating poverty rates. The study focused on five administrative departments in southwest Guatemala, indicating that CDR analysis can accurately replicate poverty estimates traditionally obtained through household surveys and censuses, especially in urban areas. The methodology involved studying call activity recorded in network operator towers, extracting consumption (incoming and outgoing calls) and mobility features (how far and how often a customer travels). They designed Voronoi polygons to represent network cell coverage and merged with the geographical units where poverty rates were surveyed. Figure 2.1 illustrates the designed Voronoi polygons using the geolocation of cell towers. The features based on activity were aggregated at the polygon network level using the customer's home location. Besides, a new set of features was included, such as the number of customers entering and residing outside the network polygon and their visitation frequency. The researchers compared CDR-based findings with World Bank

poverty estimates from national surveys and censuses to validate the accuracy of CDR analysis. Finally, the study employed machine learning techniques to explore the predictive power of CDRs in estimating poverty rates. They built a Linear Regression model to predict the poverty rates at the municipal level in Guatemala with a power of prediction of $R^2 = 0.76$, revealing that CDR data could explain a significant portion of the variance in poverty levels, especially in urban settings, where cellular penetration rates tend to be higher than in rural areas. Finally, the study suggested that CDR analysis cannot entirely replace conventional data collection methods; it can significantly enhance them by providing more frequent updates and supplementary information.

Fig. 2.1 The left image illustrates the geolocation of cell towers for a region of Guatemala, and the right image illustrates the network cell coverage for each cell tower using the Voronoi polygons. [2]

Similarly, Steele et al. [3] analyzed call detail records (CDRs) from mobile phone metadata to map poverty indicators in low- and middle-income countries. These data provide insights into social networks, call behavior, and mobility patterns, which correlate with SES. They studied mobile phone usage variables such as incoming and outgoing calls, incoming and outgoing text, percent of nocturnal calls, number and entropy of places visited, radius of gyration, interactions per contact, and entropy of contacts calculated per customer. They aggregated at a granular spatial resolution using the Voronoi polygons with the tower geolocation. They developed a Bayesian model to predict Namibia, Nepal, and Bangladesh's Health Survey (HS) wealth index. The model achieved a prediction power of $R^2 = 0.66, R^2 = 0.61$, and $R^2 = 0.64$ for the three countries. Figure 2.2 illustrates the HS wealth index prediction results for the three countries. Despite the inherent biases in CDR data, such as

skewness towards more educated, urban, and wealthier individuals, the study demonstrates that the combination of variables that measure mobile activity, geographic distances, and entropy extracted from CDRs can provide valuable information on spatial and temporal poverty variations, complementing traditional survey and census methods.

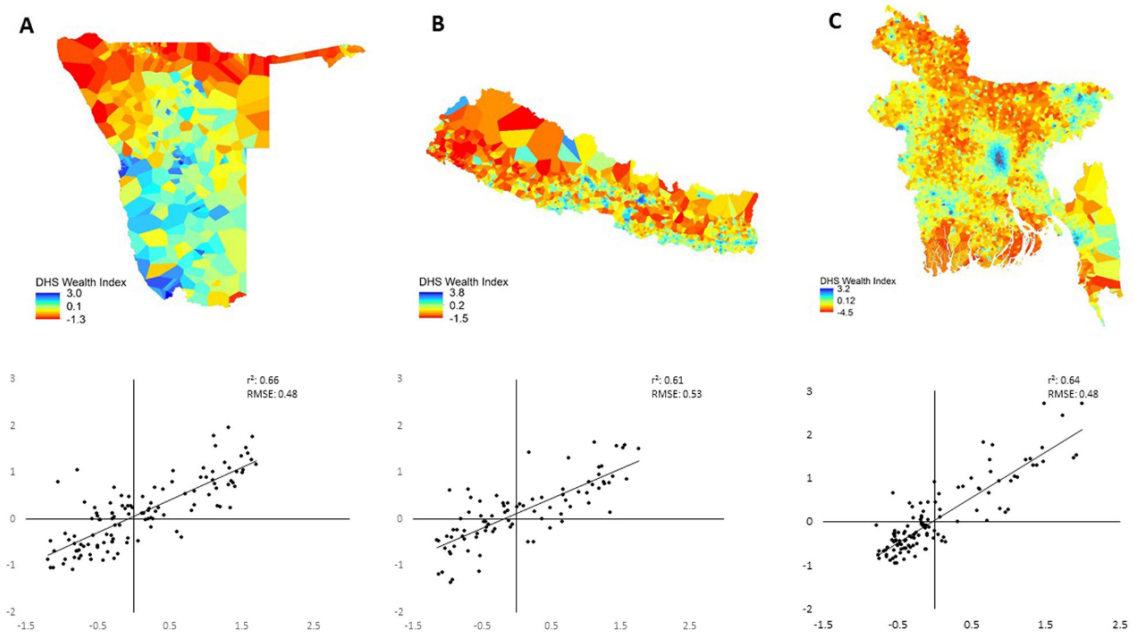


Fig. 2.2 The prediction of the HS wealth index using the Bayesian Model for (A) Namibia, (B) Nepal, and (C) Bangladesh. [3]

Despite the efforts of the scientific community to use CDRs to predict SES, these data contain sensitive information about individual's communication patterns, locations, and social networks that can raise privacy and ethical concerns. Additionally, these data are not readily available to the research community in developing countries, mainly where few mobile network operators control the market. For instance, two private mobile network operators in Ecuador control more than 83% of the market [26].

2.2 Satellite imagery to predict SES

Open access to multiple satellite imagery data sources has facilitated the execution of several studies to understand the relationship between spatial composition and poverty index at an intra-urban resolution in different regions.

Zhao et al. [4] estimated poverty levels using a Random Forest Regression (RFR) model, integrating multiple data sources to overcome limitations of existing studies that used single data types. The study focused on Bangladesh, employing data from the Demographic and

Health Surveys (DHS) to represent poverty levels through the household wealth index (WI) with a resolution of 10 km x 10km. The RFR model was trained using features extracted from various satellite imagery data sources with a resolution of 15 arc-seconds , including nighttime light data from the NPP-VIIRS Day/Night Band, Google satellite imagery, land cover, road maps, and location data of division headquarters. Figure 2.3 illustrates the datasets used in the study. The model demonstrated good predictive power, with an $R^2 = 0.70$ for Bangladesh, indicating a strong correlation between actual and estimated WI. The model was also applied to Nepal, achieving an $R^2 = 0.61$, showing its generalization capability across different geographical contexts. The study found that proximity to urban areas was the most significant variable affecting poverty estimation, contributing to 37.9% of the model's explanatory power. This highlighted the importance of accessibility and urban proximity in poverty analysis.

Similarly, Xu et al. [21] focused on poverty mapping in southwest China using multi-source geospatial data. The study employed machine learning algorithms to estimate the Integrated Poverty Index (IPI) using spatial variables such as nighttime light, traffic accessibility, altitude, flat area coverage, impervious surface coverage, cropland coverage, road density, and risk index extracted from multi-source geospatial data (NPP-VIIRS NTL, SRTM DEM, FROM-GLC, OSM) for an impoverished area of southwest China. The study compares eight machine learning algorithms and finds that the XGBoost algorithm outperforms others in estimating the IPI, with an MAE of 0.0454 and an $R^2 = 0.68$. The paper highlights the importance of multi-source data, noting that variables like proximity to urban areas significantly impact poverty levels. The results show a strong correlation between the derived IPI and SES, suggesting the potential of this method in guiding targeted poverty alleviation strategies.

Likewise, Li et al. [5] examined the application of high-resolution satellite imagery and machine learning to identify urban poverty in China, focusing on Jiangxia and Huangpi districts in Wuhan. The study employed various image features like perimeter, line segment detector (LSD), Hough transform, gray-level cooccurrence matrix (GLCM), histogram of oriented gradients (HoG), and local binary patterns (LBP) to analyze the built-up areas and their association with poverty levels. Figure 2.4 illustrates the shape and texture features extracted from images captured from high-resolution satellites to estimate the different Poverty Index (PI) levels. Four machine learning models (Random Forest, Gaussian Process Regression, Support Vector Regression, and Neural Network) were used to analyze the data, revealing that Support Vector Regression could identify Poverty Index (PI) with moderate success, with the best model achieving an $R^2 = 0.53$.

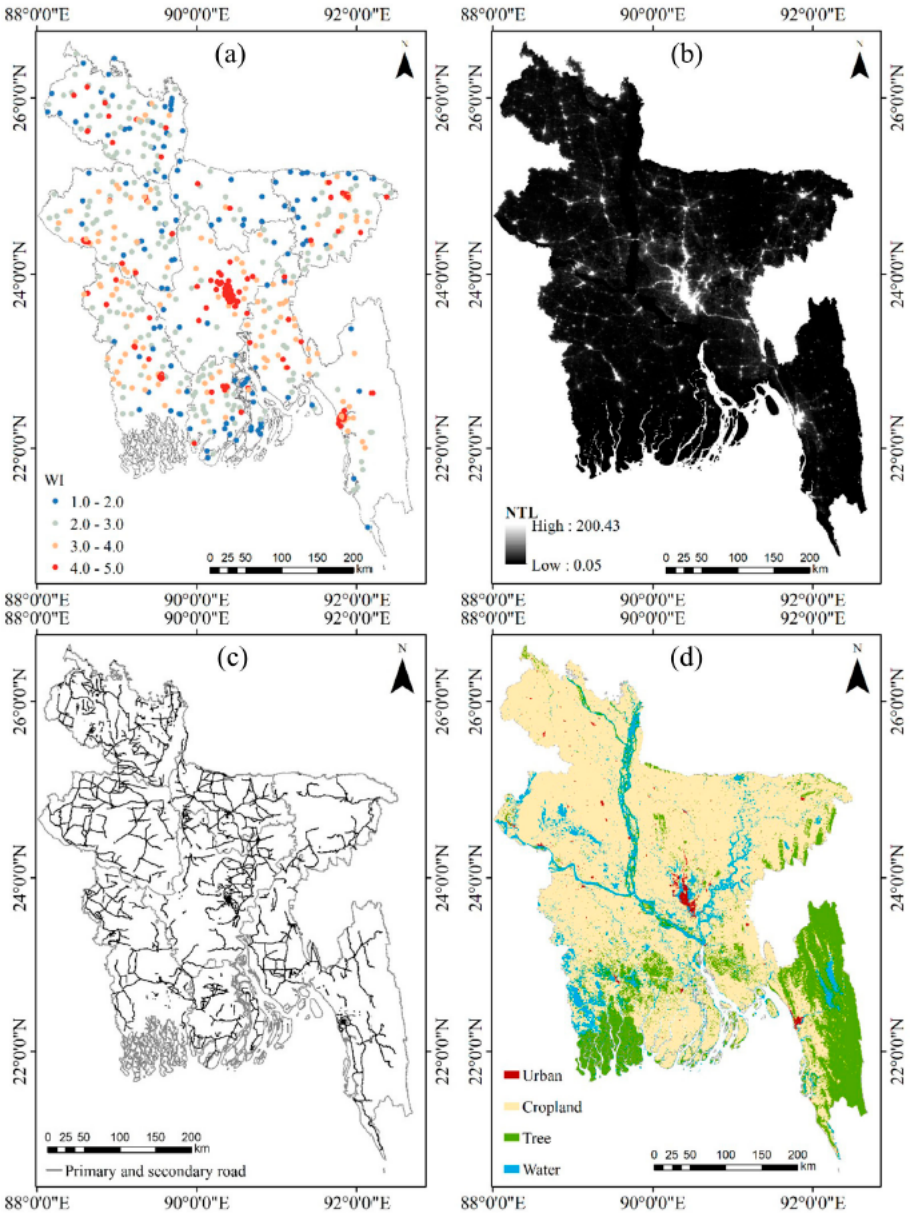


Fig. 2.3 (a) Wealth index map, (b) NPP-VIIRS nighttime light image, (c) Open street map road map, and (d) land cover map for Bangladesh. [4]

Despite the efforts of the scientific community to use satellite imagery data sources to predict SES. The studies recognize the limitations of these approaches, such as the potential biases inherent in remote sensing data to measure living conditions and the need for high-resolution imagery to capture detailed urban features, facing challenges in measuring the economic rhythms of the population.

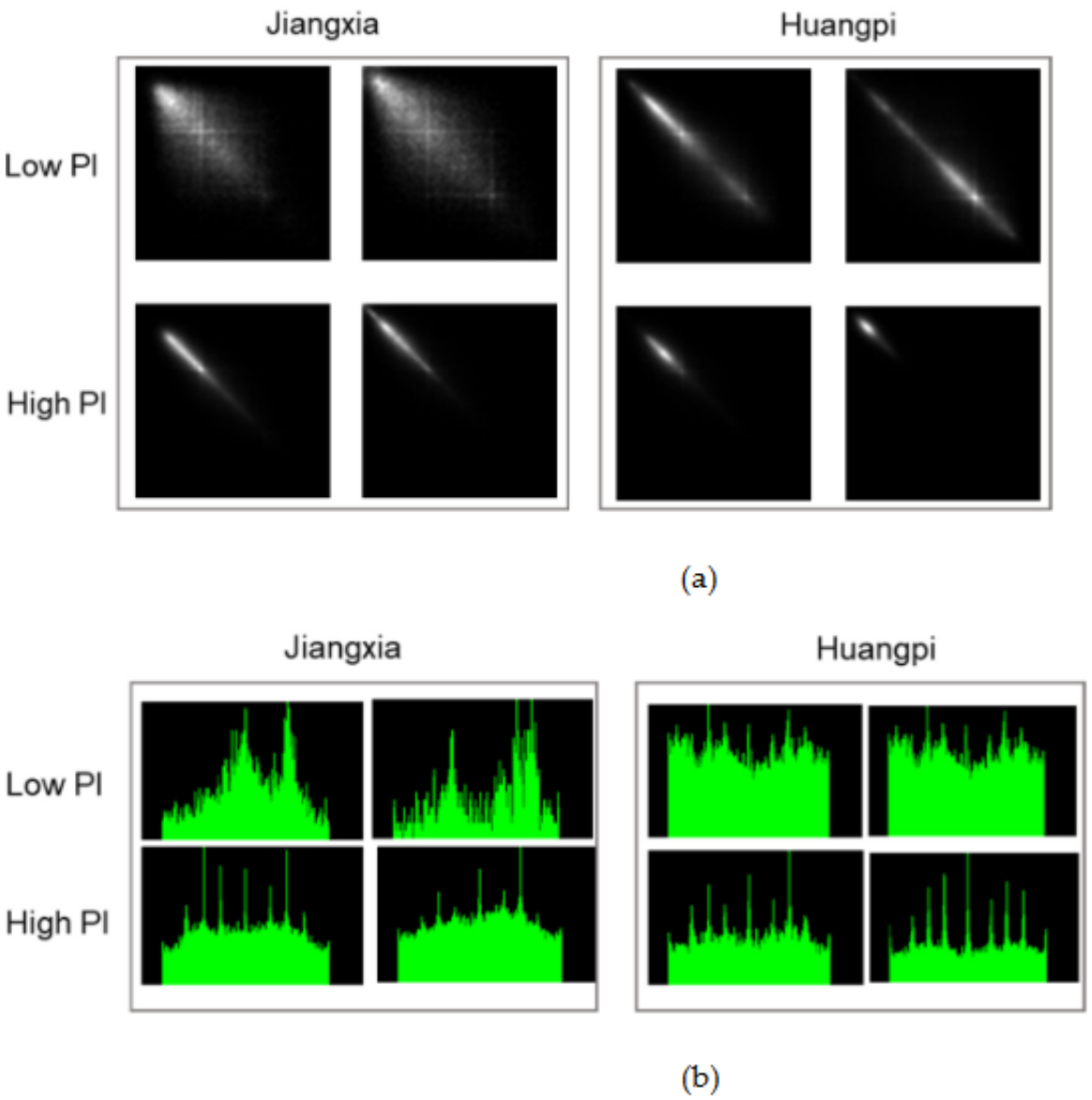


Fig. 2.4 Illustration of a shape feature (a) gray-level image representation of the GLCM, and a texture feature (b) histogram image of HoG extracted from images captured from high-resolution satellite to estimate the different levels of Poverty Index. [5]

2.3 Social media to predict SES

The growth in the use of social networks in developing countries has generated interest in studying the relationship with SES.

Giurgola et al. [6] explored the use of Facebook advertising audience estimates to analyze socioeconomic conditions in urban areas across four cities: Atlanta (USA), Bogotá (Colombia), Santiago (Chile), and Casablanca (Morocco). The study investigated the potential

of this novel data source to provide insights into the SES of populations at a high spatial granularity, particularly at the level of urban subdivisions. They performed queries using the Facebook Marketing API into a geographic area of interest with a fixed radius of 1 Km. For each query, they requested the number of users who matched over 47 attributes associated with gender, marital status, education, travel, interests, religion, technology, and connectivity. Figure 2.5 illustrates the spatial targeting schemes for the Facebook Marketing API queries of Bogota, Santiago, and Casablanca. They performed a Linear Regression model to predict median household income, socioeconomic strata, and multidimensional poverty rate. The model reached an $R^2 = 0.93$ for Santiago, while for the other three cities was between $R^2 = 0.46$ and $R^2 = 0.56$ with users older than 25.

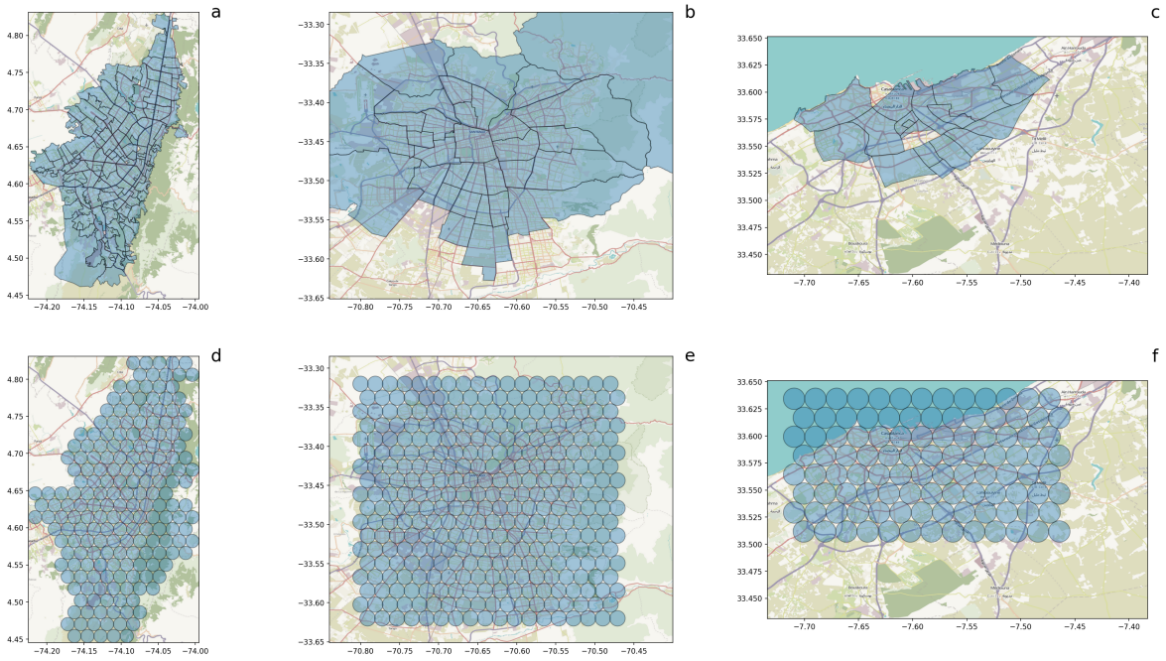


Fig. 2.5 The first row illustrates the administrative subdivisions for Bogota (a), Santiago (b), and Casablanca (c). The second row illustrates the spatial targeting schemes of a radius of 1 Km for the Facebook Marketing API queries of Bogota (d), Santiago (e), and Casablanca (f). [6]

However, the update process of estimating SES using Facebook marketing data is limited by the update dynamics, executed based on internal programming and the company's needs. Furthermore, it is questioned whether Facebook algorithms for inferring the categories of fields that users have not completed and are used for estimating SES work adequately across different regions and populations [6].

2.4 Graph Neural Networks to predict SES

The neural networks can model complex, non-linear relationships due to their multiple layers and non-linear activation functions. They are suitable for tasks where the relationship between features and the target is highly non-linear. Recent studies have used graphs as data structures to model connections and, combined with neural networks, have formed architectures to predict poverty.

Ma et al. [7] explored the challenge of identifying poverty at the village level, proposing a novel method utilizing graph-based modeling to understand and predict poverty status. The study is set in Enshi, a poverty-stricken area in China, where the authors collect data to build a village graph representing the geographic and economic relationships between villages. They introduced two main concepts: Centrality and Homophily Decaying effect. Centrality refers to the position of a village within the graph, indicating its connectivity and potential economic activity. The Homophily Decaying effect observes that villages closer to each other tend to have similar poverty statuses, but this similarity decreases as the distance increases. Moreover, they employed a Global Centrality2Vec module to embed the centrality features into a dense vector and a Local Graph Distance Convolution module to account for the decaying similarity effect over distance. Figure 2.6 shows the model framework to predict poverty using centrality and homophily decaying effect. These methods aim to capture the complex interplay between geographic proximity and economic activity in determining poverty. The model achieved an accuracy of 0.76 and an F1 score of 0.54 for predicting the poverty level at the village level.

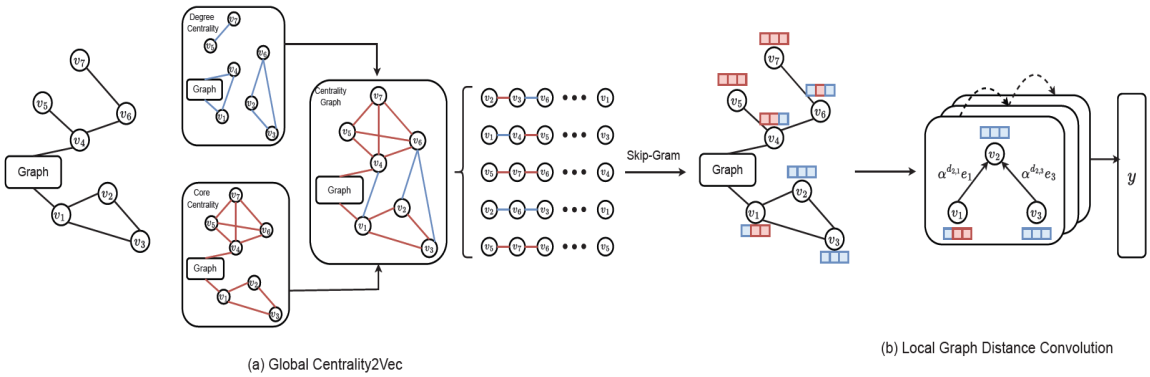


Fig. 2.6 Model framework to predict poverty using centrality and homophily decaying effect. [7]

Likewise, Cao et al. [8] introduced a machine-learning approach to estimate and map economic development using multi-source open geospatial data. The method leverages nighttime light imagery, multispectral remote sensing imagery, and OpenStreetMap road

networks to extract knowledge-based features. These features are then processed through a multi-view graph neural network (MVGNN) to predict regional economic indicators, such as Gross Domestic Product (GDP), in a self-supervised learning manner. Figure 2.7 illustrates the model architecture of a Multi-view graph neural network (MVGNN). Experiments conducted across China's counties demonstrate the method's effectiveness in estimating GDP, with the integrated approach of knowledge-based and learning-based features significantly outperforming baseline methods. The proposed model reached a performance of $R^2 = 0.82$. The research highlights the potential of open geospatial data in providing accurate, timely SES for smart governance and policy-making.

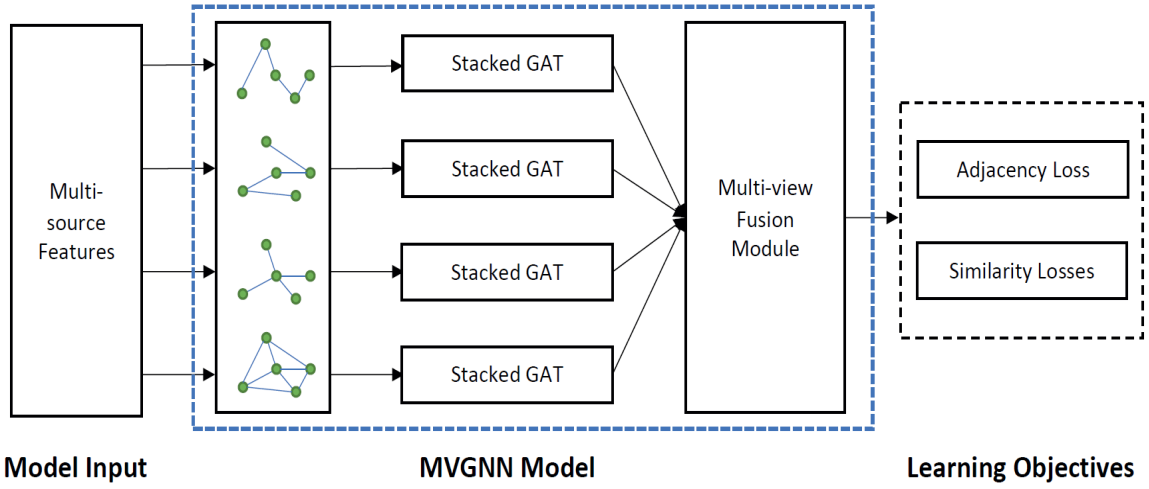


Fig. 2.7 Model architecture of Multi-view graph neural network (MVGNN). [8]

However, These models primarily use static geospatial features. They may not fully capture the dynamic nature of economic development, which can be influenced by a wide range of temporal factors and events not represented in the static geospatial data.

2.5 Summary of recent studies to predict SES

Several studies have analyzed surveys, CDRs, satellite imagery, geographic information, and social media, providing valuable information about predicting SES to overcome traditional method collection problems at the intra-urban level. Table 2.1 describes a summary of recent studies on predicting SES using digital sources. On the one hand, several studies described in this literature review have employed linear regression and decision tree-based models to capture the linear relationship between the features extracted from digital data to predict a continuous SES. On the other hand, recent studies have employed neural networks to

model complex non-linear relationships between data structures as graphs and the SES using multiple layers and non-linear activation functions. However, with the availability of other digital data sources that record the purchasing activity of customers living in an urban neighborhood, there is great potential in exploring and exploiting machine learning and deep learning models to learn about the linear and non-linear relationships between the purchase behavior and the SES at the neighborhood level to improve the predictive power, which will be beneficial for measuring the economic rhythms of the population.

Table 2.1 Summary of recent studies that have worked on predicting SES using digital data sources.

Ref	Year	Country	Data	Method	SES Index	Power of prediction
[2]	2017	Guatemala	CDRs	Linear Regression	World Bank poverty	$R^2 = 0.76$
[3]	2021	Namibia, Nepal, and Bangladesh	CDRs	Linear Regression	Health Survey (HS) wealth index	$R^2 = 0.66$, $R^2 = 0.61$, $R^2 = 0.64$
[4]	2019	Bangladesh, Nepal	Nighttime light, Google images, road map, land cover map, division headquarters location data	Random Forest Regression	Household Wealth Index (WI)	$R^2 = 0.70$, $R^2 = 0.61$

[21]	2021	China (southwest)	Official state publications, accessibility to cities, average nighttime light, SRTM/DEM, land cover, and natural disaster data	XGBoost	Integrated Poverty Index (IPI)	$R^2 = 0.68$
[5]	2021	China (Jiangxia and Huangpi)	Line segment detector (LSD), Hough transform, gray-level cooccurrence matrix (GLCM), histogram of oriented gradients (HoG), and local binary patterns (LBP)	Support Vector Regression	Poverty Index (PI)	$R^2 = 0.53$

[6]	2021	Atlanta, Bogotá, Santiago, and Casablanca	Facebook Marketing API	Linear Regression	Household income, Socioeconomic strata, and Multidimensional poverty rate	$R^2 = 055$, $R^2 = 0.56$, $R^2 = 0.93$, $R^2 = 0.46$
[7]	2021	China	Oficial geographic data	Graph Neural Network	Poverty level	$Acc = 0.76$
[8]	2022	China	Nighttime light imagery, multispectral remote sensing imagery, Open Street Map, and GDP	Multi View Graph Neural Network	Gross Domestic Product	$R^2 = 0.82$

Chapter 3

Datasets

NSIs of developing countries measure the socioeconomic status using censuses and surveys, which have the following disadvantages: long-time intervals, limited population coverage, and the results are reported with coarse spatial granularity, generally at the country or city level [2, 27, 6].

This chapter examines the alternative digital data sources employed in this work as a proxy for estimating socioeconomic status at the intra-urban level to measure the economic rhythms of the population, enabling the analysis of socioeconomic status at a more fine spatial resolution that is difficult to achieve by traditional methods of collecting socioeconomic status that are costly in time and resources. This study is focused on Ecuador, and the alternative digital data sources collected include:

- Urban neighborhoods, described in section 3.1.
- Top-up transactions, described in section 3.2.
- Supermarket transactions, described in section 3.3.
- Neighborhood Socioeconomic Status, described in section 3.4.

3.1 Urban neighborhoods

Ecuador is geographically distributed in administrative boundaries to organize public management in a decentralized and deconcentrated manner. The administrative boundaries are divided into three levels: level 1 (provinces), 2 (cantons), and 3 (parishes). The geographic maps are publicly available in digital format. However, the boundaries of the third administrative level are of coarse spatial resolution. They may contain several urban

neighborhoods with different socioeconomic characteristics [17, 15, 16], and this study is conducted at the urban neighborhood level, a fine spatial resolution intra-urban, which is defined as a residential geographic area of a more granular level (less than 10 square kilometers) where most households share similar socioeconomic characteristics.

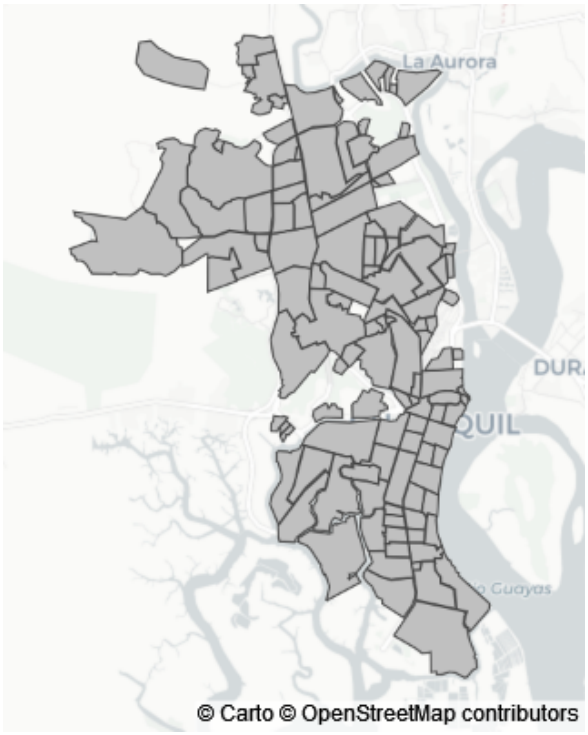
The official boundaries of the urban neighborhoods are not publicly available in digital format. Therefore, a collection process was performed to replicate the urban neighborhood boundaries of Guayaquil and Quito (Ecuador's two most populated cities) using different geographic information sources such as Google Maps, Nominatim with OpenStreetMap, and Wikimapia. The result was a geographic dataset in Geo-JSON format, containing 97 and 46 neighborhoods for Guayaquil and Quito, respectively, with an area ranging from 2 to 9 square kilometers. Figure 3.1 illustrates the urban neighborhoods of Guayaquil and Quito.

3.2 Top-up transactions

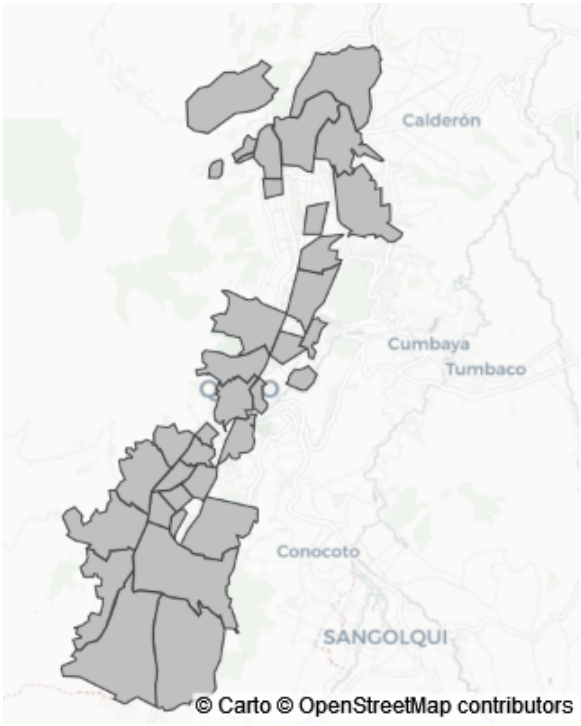
The International Telecommunications Union (ITU) reported that the number of cellular mobile phone subscriptions in Latin America was between 105 and 109 per 100 people from 2018 to 2022 [1] [2]. In most Latin American countries, mobile network operators offer two subscription options for their mobile phone customers: postpaid and prepaid. In the postpaid subscription, a customer is committed to a long-term contract with the operator, and the service is billed at the end of each month. On the other hand, in the prepaid subscription, credit is purchased in advance (**top-up transaction**) by the customer, and the use of the service is granted only if there is available credit in the mobile line. Unlike developed countries, the prepaid subscription is the preferred one in Latin America countries [28, 29], and prepaid packages are used by most people in the region when buying mobile data bundles for affordable access to the Internet [30]. For instance, in Ecuador, the density of active mobile phone lines in the country was 93.44%, with 78.20% being prepaid subscriptions in September 2021 [3]. Consequently, analyzing mobile airtime/data package purchasing dynamics can help derive proxies for estimating neighborhood socioeconomic status since the associated data includes spatial data reflecting when and where people have purchased.

Ecuador currently has three wholesale suppliers that process top-up transactions of all mobile network operators through points of sale located in grocery stores, pharmacies, supermarkets, etc. The dominant provider produced an anonymized dataset with over 9M top-up transactions in 2017. The dataset recorded 788,701 unique customers who performed

[1] https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=ZJ
[2] https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx
[3] https://www.arcotel.gob.ec/servicio-movil-avanzado-sma_3/



(a) Urban neighborhoods of Guayaquil



(b) Urban neighborhoods of Quito

Fig. 3.1 Urban neighborhoods of Guayaquil (a) and Quito (b).

top-up transactions in 3,584 unique shops geographically distributed across Guayaquil and Quito. Each top-up transaction recorded the timestamp, the customer ID, the shop ID, and the total amount of purchased credit. Table 3.1 shows the format of the top-up transaction records. Moreover, a complementary dataset was provided with a unique ID for each shop involved in the transactions and its corresponding geographic coordinates. Each shop has a pair of latitude/longitude associated with it, and those points are distributed across the entire city.

The company removed sensitive information by transforming the customer ID into an anonymized representation since it corresponds to the telephone number to which the credit is purchased. Moreover, the company applied an additional level of anonymization to change the total amount of credit by using a function that preserves the original ranking.

Table 3.1 (a) The format of top-up transactions and (b) the shop's location

(a)

timestamp	customer ID	shop ID	total amount of credit
2017-03-05 09:15	%¡¿?=*+	56742	0.23
2017-03-06 10:14	%¿¡¬l]'	987623	0.17
.	.	.	.
.	.	.	.

(b)

shop ID	latitude	longitude
56742	-2.18	-79.90
987623	-2.11	-79.79
.	.	.
.	.	.

3.3 Supermarket transactions

Three supermarket chains with nationwide coverage dominate Ecuador's convenience store market. These chains offer a range of products, including groceries, personal hygiene products, and cleaning supplies. One of these supermarket chains allowed access to anonymized purchase data from its 257 stores. Our study specifically focused on the 53 stores in Guayaquil, where this chain has its highest concentration of stores.

In Guayaquil, the chain recorded over 20M shopping transactions from January 1, 2016, to December 31, 2018. Each transaction contains the date, transaction ID, customer ID, item ID, quantity, and store ID (refer to Table 3.2). Additionally, the supermarket chain provided two complementary datasets:

1. A dataset in which each store is linked to geographic coordinates (latitude and longitude), refer to Table 3.2.

2. A dataset in which each item is associated with its family and price, refer to Table 3.2.

It is important to note that the supermarket chain excluded sensitive customer information from the datasets.

Table 3.2 (a) The format of supermarket transactions, (b) the format of geographic information for the stores, and (c) the format of the features for the items.

(a)

Date	Transaction ID	Customer ID	Item ID	Qty	Store ID
2016-01-01	1001	101	101	10	1
2016-01-01	1001	101	202	5	1
2016-01-01	1002	902	402	8	2
2016-01-01	1002	902	602	3	2
.	.	.	.	.	.
.	.	.	.	.	.

(b)

Store ID	Latitude	Longitude
1	-2.25754	-79.89032
2	-2.27512	-79.87929
.	.	.
.	.	.

(c)

Item ID	Item Family	Unit Price
101	CANNED FISH	2.45
202	MILK CARTON	1.10
402	SODA	0.56
602	SAUSAGES	0.47
.	.	.
.	.	.

3.4 Neighborhood Socioeconomic Status

Since 1993, the NSI of Ecuador (Instituto Nacional de Estadísticas y Censos (INEC)) has conducted the National Survey of Employment, Unemployment, and Underemployment (ENEMDU). This survey, which is carried out quarterly, provides national coverage and explores various aspects such as the employment situation, labor market characterization,

economic activities, and income sources of the population. INEC utilizes a spatially disaggregated sampling frame based on census tracts, typically averaging 200 square meters with around 150 dwellings each, for sociodemographic research. Although the ENEMDU survey delivers country-level results, its coverage is constrained to sampling census tracts across all provinces due to significant demands on resources and time, as noted in reference [31]. This limitation means that the survey covers only certain urban neighborhoods in cities, posing challenges for analyzing the socioeconomic realities of areas not included in the surveyed tracts. For instance, Figure 1.2 shows the surveyed census tracts and those urban neighborhoods in Guayaquil where socioeconomic indicators were unavailable in the first quarter of 2017.

From 2016 to 2018, the socioeconomic survey data was collected for both Guayaquil and Quito, corresponding with the period of top-up and supermarket transactions. To calculate the socioeconomic status of each neighborhood, these steps were followed: First, a spatial join was executed between the surveyed census tracts and the urban neighborhood boundaries. Subsequently, we calculated the average per capita income from ENEMDU, which was then designated as the Neighborhood Socioeconomic Status (NSES). This NSES served as the ground truth in our supervised inference method.

Chapter 4

Methodology

This chapter outlines the architecture and components of two distinct machine learning models: Regression and Graph Neural Network (GNN) models. This study proposes these models to predict Neighborhood Socioeconomic Status (NSES) using data from top-up and supermarket transactions. These models are selected to address specific aspects of the socioeconomic estimation problem where traditional statistical methods may fall short due to the sparse and networked nature of the available data.

The Regression Model is designed to capture linear relationships between known predictors of consumption behavior derived from the top-up transactions dataset (described in section 3.2) and the NSES. This model is particularly suited for quantifying the direct impact of observed economic activities on socioeconomic status. On the other hand, the Graph Neural Network model leverages the inherent network structure of supermarket transactions (described in section 3.3) to capture complex, non-linear interactions between entities, providing a more nuanced understanding of socioeconomic patterns.

4.1 Regression Model using top-up transactions

The proposed methodology systematically builds a Regression Model that utilizes consumption behavior indicators from the top-up transactions dataset described in section 3.2 to estimate Neighborhood Socioeconomic Status (NSES). The development process involves several phases, including:

1. Reducing sparsity in the dataset to improve model accuracy.
2. Detecting home locations to anchor the socioeconomic data geographically.
3. Calculating indicators of consumption behavior to serve as model inputs.

4. Spatially aggregating these indicators to match the granularity of NSES.

5. Building the Regression Model to relate these indicators with socioeconomic outcomes.

Figure 4.1 illustrates the architecture of the proposed methodology. Next, each phase is explained in detail.

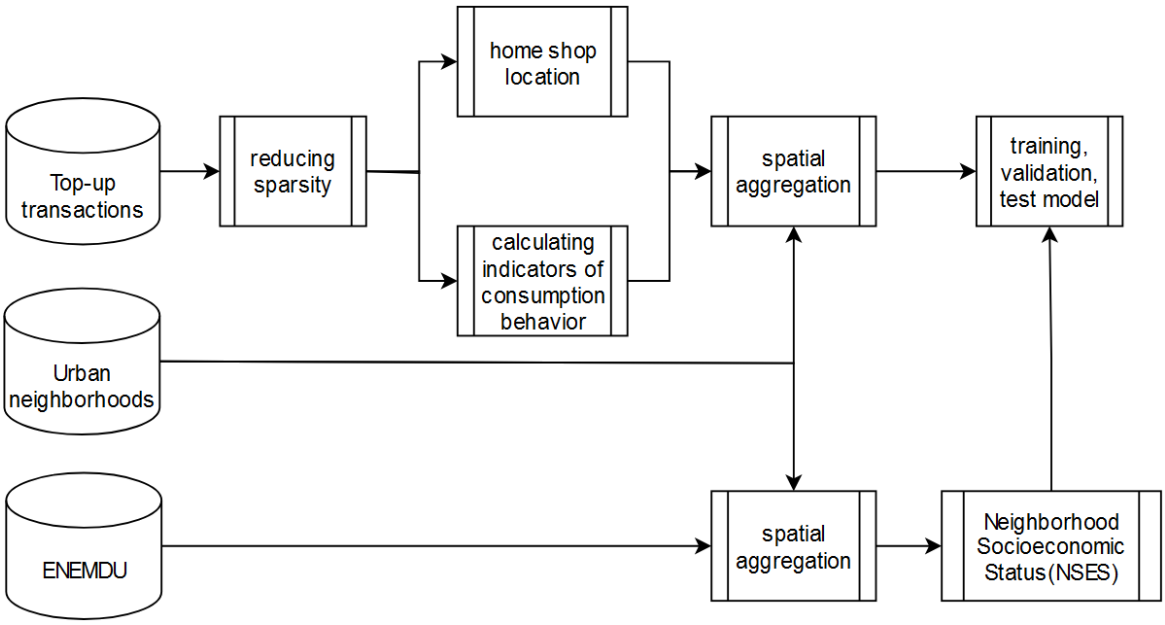


Fig. 4.1 The architecture of the proposed methodology to predict the Neighborhood Socioeconomic Status (NSES) using top-up transactions.

4.1.1 Reducing sparsity

A sparsity reduction process was performed using the top-up transactions dataset, selecting customers whose top-up transactions will be part of the study. To facilitate this process, the number of top-up transactions per year was calculated for each customer. Customers who averaged more than one transaction per month over a period of six months were selected. This criterion helped eliminate customers whose top-up transactions occurred sporadically, ensuring the data represented consistent customer behavior.

4.1.2 Home Shop location

To spatially aggregate customers to an urban neighborhood, a home location for each was determined based on the shop most frequently visited under certain conditions. The calculation excluded shops located in non-residential areas, such as educational institutions,

bus stations, gas stations, and shopping malls, because transactions at these locations are unlikely to reflect residential proximity. The shop most frequented by each customer c during working days from 7 PM to midnight was designated as the home shop $S_h(c)$ for the customer. This method successfully inferred the home locations for 440,705 unique customers.

4.1.3 Calculating indicators of consumption behavior

Top-up transactions were analyzed quarterly, aligned with the periods when NSES was assessed, to develop meaningful features that characterize consumption behavior and economic activity. The following features were calculated for each customer c.

- **Average denomination**: the average denomination of airtime top-up transactions. A customer chooses whether to top-up 0.09, 0.10, 0.11 units, or any other denomination (remember that the company has linearly transformed the monetary value involved in the transaction). A higher value of the average denomination would be an indicator of better socioeconomic status [16, 32].

$$Ad_c = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{4.1}$$

- **Denomination diversity**: the diversity of denominations the customer uses when purchasing mobile airtime. We measure the denomination diversity of the customer c by using the Shannon entropy.

$$Dd_c = \sum_{i=1}^{n} p(d_i)\log p(d_i) \tag{4.2}$$

  where n is the number of denominations used by the customer c and $p(d_i)$ is the probability that the customer c uses the denomination d_i. A higher value of denomination diversity means that the customer uses different denominations to purchase mobile airtime. Previous studies have employed various diversity measures [15, 17], such as network diversity [33], travel diversity[34, 35], to quantify important aspects of socioeconomic status using CDRs.

- **Shop diversity**: the diversity of shops the customer visits.

$$Sd_c = \sum_{i=1}^{n} p(s_i)\log p(s_i) \tag{4.3}$$

Table 4.1 Pearson correlation between all predictors and the per capita income.

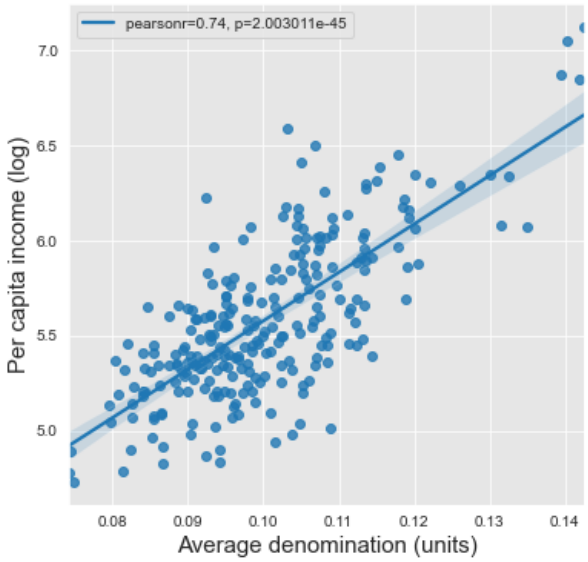
Features	ρ	p value
average denomination	0.74	2.00e-45
denomination diversity	0.55	3.04e-22
shop diversity	0.46	3.92e-15

where n is the number of shops visited by the customer c and $p(s_i)$ is the probability that the customer c visits the shop s_i. A higher value of shop diversity means customers diversify their purchases in different stores.

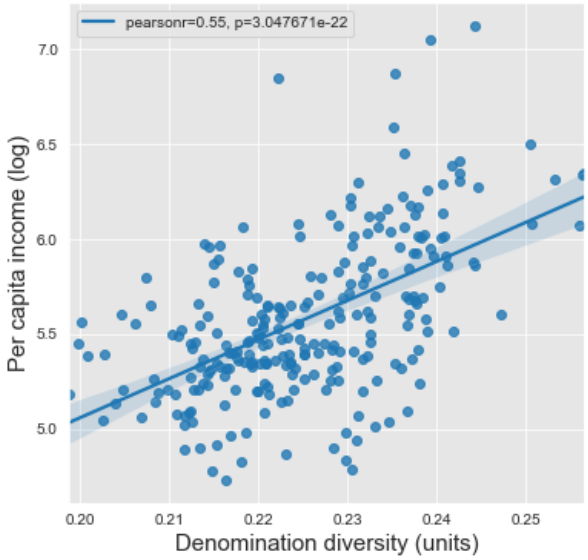
4.1.4 Spatial aggregation

The spatial aggregation process began by assigning each customer c to an urban neighborhood based on the longitude and latitude of their most frequented shop, $S_h(c)$. We then calculated the distribution of the customer population across all urban neighborhoods. Urban neighborhoods with a population density below the first quartile were excluded from further analysis. This resulted in 68 neighborhoods with an average of 1,352,183 top-up transactions and 162,665 per quarter in Guayaquil and Quito in 2017. Next, we aggregated customers' consumption behavior indicators to their respective urban neighborhoods using a statistical bootstrap to calculate the average. This produced a real-value vector representing residents' economic behavior within each neighborhood. Finally, we combined all the neighborhood vectors from the four quarters to create a dataset comprising 278 neighborhoods. This dataset was used to train a prediction model to estimate the NSES.

The correlation between the consumption behavior indicators and the NSES was calculated across all the urban neighborhoods. Table 4.1 displays the Pearson correlation coefficients for the three indicators of consumption behavior and NSES. Figure 4.2 illustrates these correlations. The average denomination shows the strongest positive correlation $\rho = 0.74$, indicating a linear correlation with the NSES, suggesting that customers who reside in wealthier urban neighborhoods tend to make higher-value top-up transactions. Figure 4.3 presents a choropleth map highlighting the spatial correlation between the average denomination and per capita income, where urban neighborhoods with a higher average denomination are predominantly wealthier. Additionally, denomination diversity exhibited a notable positive correlation $\rho = 0.55$, suggesting that residents of wealthier neighborhoods tend to top-up mobile airtime with diverse denominations. Finally, the shop diversity showed a moderate positive correlation $\rho = 0.46$, implying that customers in wealthier areas tend to distribute their top-up transactions across different shops.

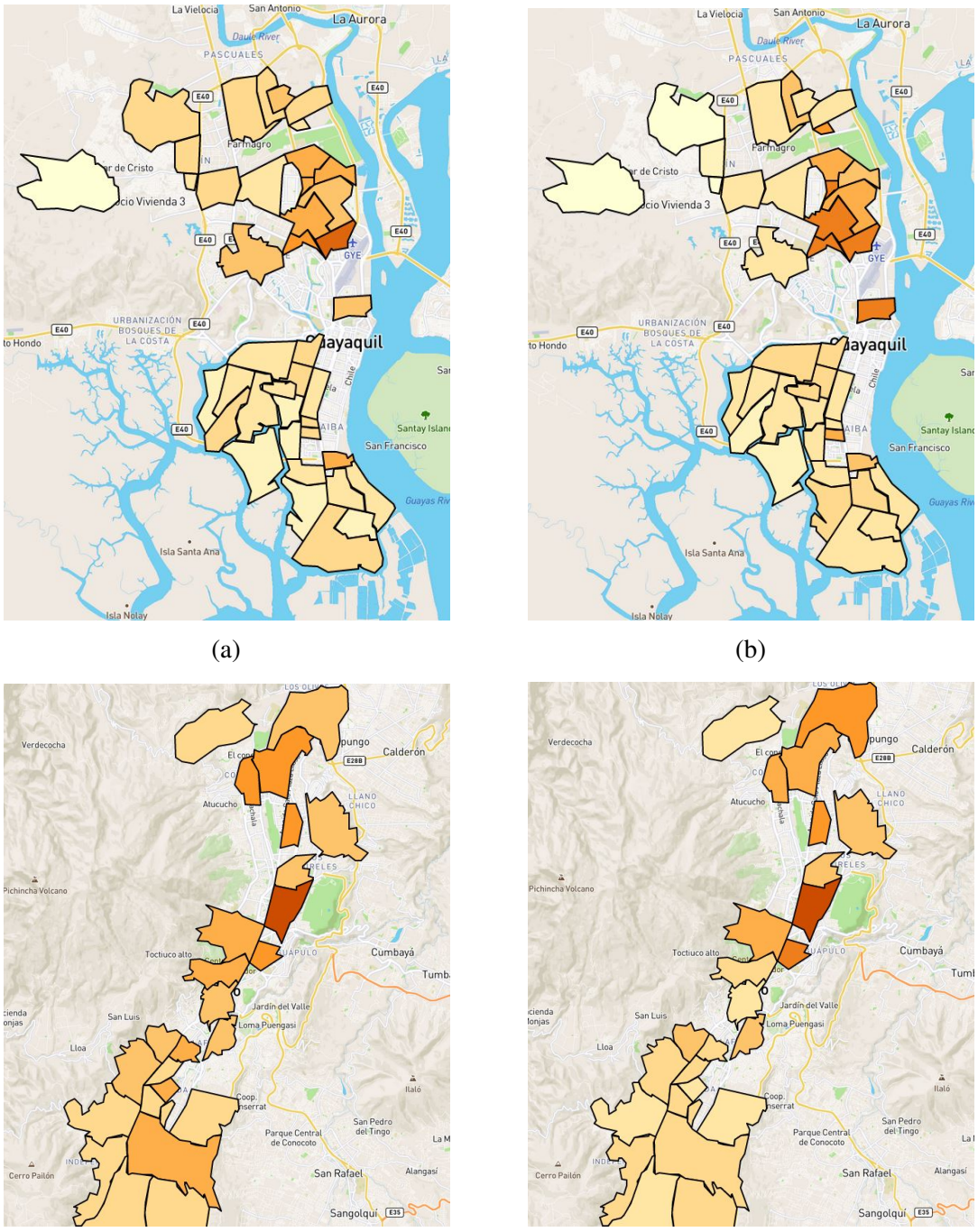


(a) Correlation between average denomination (units) and Neighborhood Socioeconomic Status (log).



(b) Correlation between denomination diversity (units) and Neighborhood Socioeconomic Status (log).

Fig. 4.2 Correlation between indicators of consumption behavior and Neighborhood Socioeconomic Status.



(a) (b)

(c) (d)

Fig. 4.3 Choropleth maps showing the relationship between the average denomination and the average per capita income (NSES). The first column illustrates the spatial distribution of the average denomination for the urban neighborhoods of (a) Guayaquil, and (c) Quito. The second column illustrates the spatial distribution of the average per capita income (NSES) for the urban neighborhoods of (b) Guayaquil, and (d) Quito.

4.1.5 Building the Regression Model

The prediction of the NSES was framed as a regression task, with indicators of customers' consumption behavior serving as the independent variables. The neighborhood dataset was split into a training set (70%, 194 neighborhood vectors) and a testing set (30%, 84 neighborhood vectors). During the training phase, we implemented a feature selection mechanism using five-fold cross-validation to identify the most predictive features. Five machine-learning models were evaluated to determine the best model for predicting NSES: Linear Regression, Lasso (Linear Regression with L1 regularization), Ridge (Linear Regression with L2 regularization), Random Forest Regression, and Support Vector Regression.

4.2 Graph Neural Network Model using supermarket transactions

The proposed methodology is a systematic process that transforms supermarket transactions described in section 3.3 into a graph representation for a deep learning model to predict the socioeconomic status at the urban neighborhood level (NSES). The process involves three phases:

1. The creation of an item embedding $\hat{I}$ based on their family and price.
2. The representation of a Basket Graph G using transaction data from urban neighborhood stores to represent item co-purchases.
3. The construction of a Graph Neural Network to predict Neighborhood Socioeconomic Status (NSES) using the Basket Graph G.

Figure 4.5 illustrates the architecture of the proposed methodology. Next, each phase is explained in detail.

4.2.1 Item Embedding

A module to encode item features as a fixed-size embedding was developed and utilized as node attributes in the graph representation. Given $F = \{f_1, f_2, \ldots, f_i\}$ representing the set of item families, a Doc2Vec model was trained [36] to encode each item family in a real-valued vector of size 8 using the set of item families as the corpus, producing a new set of item family embeddings $\hat{F} = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_i\}$. Additionally, each item family $f_i \in \hat{F}$ has a set of prices

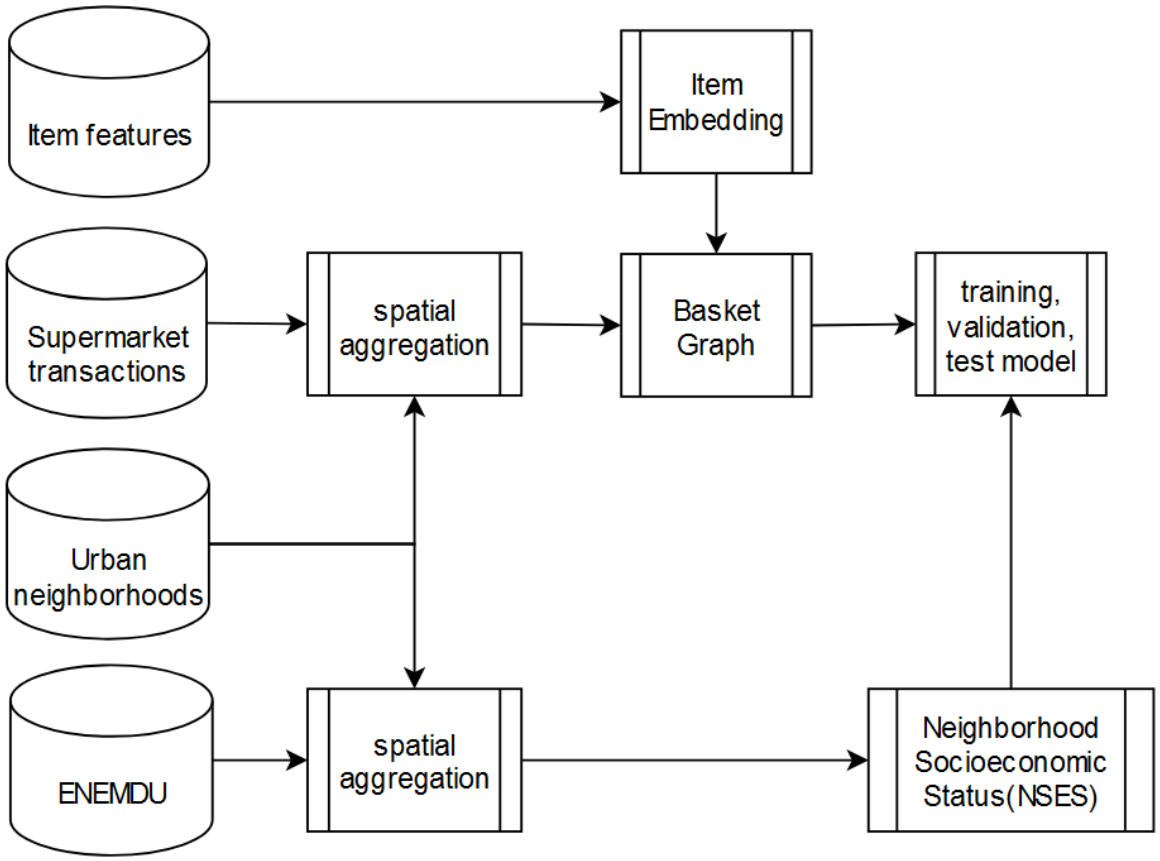


Fig. 4.4 The architecture of the proposed methodology to predict Neighborhood Socioeconomic Status (NSES) using supermarket transactions.

$P_{fi} = \{p_1, p_2, \ldots, p_j\}$. Such prices were standardized using MinMaxScaler with a range between 0.1 and 0.9 to accommodate the totals where the price is multiplied by quantity. Such a process produced a new set of standardized item prices $\bar{P_{fi}} = \{\bar{p_1}, \bar{p_2}, \ldots, \bar{p_j}\}$ for each item family f_i. Therefore, an item embedding combines the family embedding and the standardized price, producing a real-valued set of item embeddings $\hat{I} = \{(\hat{f_i}, \bar{p_j}) \mid \hat{f_i} \in \hat{F}, \bar{p_j} \in \bar{P_{fi}}\}$, where $\hat{I} \in \mathbb{R}^d$, and $d = 9$.

4.2.2 Spatial aggregation

The catchment area of each supermarket was calculated using a projection radius of 500 meters around the supermarket's geolocation point and spatially intersected with the polygons of the urban neighborhoods. The polygon with the largest intersection area was assigned as the supermarket´s urban neighborhood. The 53 supermarkets in Guayaquil were spatially joined into 29 urban neighborhoods.

4.2.3 Basket Graph

These supermarket transactions were then transformed into a graph representation called Basket Graph [37] to model relationships between purchased items and understand customers' consumption behavior within an urban neighborhood. Given $T_n(r) = \{t_1, t_2, \ldots, t_m\}$ as the set of supermarket transactions for the neighborhood n during the quarter r; where $t_m = \{((\hat{f}_1, \bar{p}_1), q_1), ((\hat{f}_2, \bar{p}_1), q_2), \ldots, ((\hat{f}_i, \bar{p}_j), q_k)\}$ was defined as the set of item embeddings $(\hat{f}_i, \bar{p}_j) \in \hat{I}$ with their quantities $q_1, q_2, \ldots, q_k$ purchased in the transaction t_m. The Basket Graph was defined as a weighted attributed graph $G = (V, E, W)$, where $V \in \mathbb{R}^{|V| \times d}$ are the attributed nodes that belong to the set of item embeddings $\hat{I}$ that were purchased in the supermarket stores for the urban neighborhood n, $E \in \mathbb{R}^{|E|}$ denotes the set of links representing pairs of item embeddings purchased within the same transaction t_k, and $W \in \mathbb{R}^{|E| \times 1}$ represents the set of weights for each link in E. The weight $w_{xy} \in W$ for the link between the item embedding x and y was calculated using the following equation:

$$w_{xy} = \sum_{s=1}^{S} \frac{(\bar{p}_x * q_x) + (\bar{p}_y * q_y)}{s} \tag{4.4}$$

For the link between item embeddings, x and y were calculated based on the number of transactions S in which x and y were purchased together, p is the standardized price, and q is the purchased quantity. Links with weights below the third quartile were removed to exclude random item co-purchases. This process resulted in 232 Basket Graphs representing customer consumption behavior across the urban neighborhoods for the observed quarters.

4.2.4 Graph Neural Network Model

Graph Neural Network (GNN) models can be built using two main approaches to define convolutional filters using either a spatial approach, which localizes filters to operate within a finite-sized neighborhood of nodes, to understand node properties based on local interactions, and the spectral approach which involves eigendecomposition of the Laplacian matrix to understand its underlying structure [38]. This study focuses on leveraging the Basket Graph's structure, representing customers' consumption behavior, to predict the neighborhood socioeconomic status (NSES). A GNN model was developed using the spectral approach with the Basket Graph as input to achieve this. It employs three convolutional layers with Chebyshev filters of size 4 ($K = 4$) and 64 channels to learn the graph's intricate patterns. Each layer in this setup utilizes the following equation:

$$\mathbf{X}' = \sigma(\sum_{k=1}^{K} \mathbf{Z}^{(k)} \cdot \theta^{(k)}) \tag{4.5}$$

$$\begin{aligned} \mathbf{Z}^{(1)} &= \mathbf{X} \\ \mathbf{Z}^{(2)} &= \hat{\mathbf{L}} \cdot \mathbf{X} \\ \mathbf{Z}^{(k)} &= 2 \cdot \hat{\mathbf{L}} \cdot \mathbf{Z}^{(k-1)} - \mathbf{Z}^{(k-2)} \end{aligned} \tag{4.6}$$

where σ is a nonlinear function (Relu), K is the Chebyshev filter size; $\mathbf{Z}^{(k)}$ is the Chebyshev polynomial of order k computed recursively and evaluated at the normalized Laplacian $\hat{\mathbf{L}}$ (a diagonal matrix of scaled eigenvalues) $\hat{\mathbf{L}} = \frac{2\mathbf{L}}{\lambda \mathbf{max}} - \mathbf{I}$; and $\theta^{(k)}$ is a vector of learnable polynomial coefficients [38]. A dropout of 0.25 was applied after each convolutional layer to prevent overfitting. An average pooling function aggregated the graph data and connected it with a linear layer to construct the predicted NSES. The 360 basket graphs from 2016 to 2017 are used to train the GNN model using Mean Square Error as the learning objective because NSES is a continuous variable, and the 103 basket graphs from 2018 are used to test the GNN model. The GNN model was implemented using PyTorch Geometric [39]. Figure 5.1 illustrates the architecture of the Graph Neural Network (GNN) model to predict the neighborhood socioeconomic status (NSES).

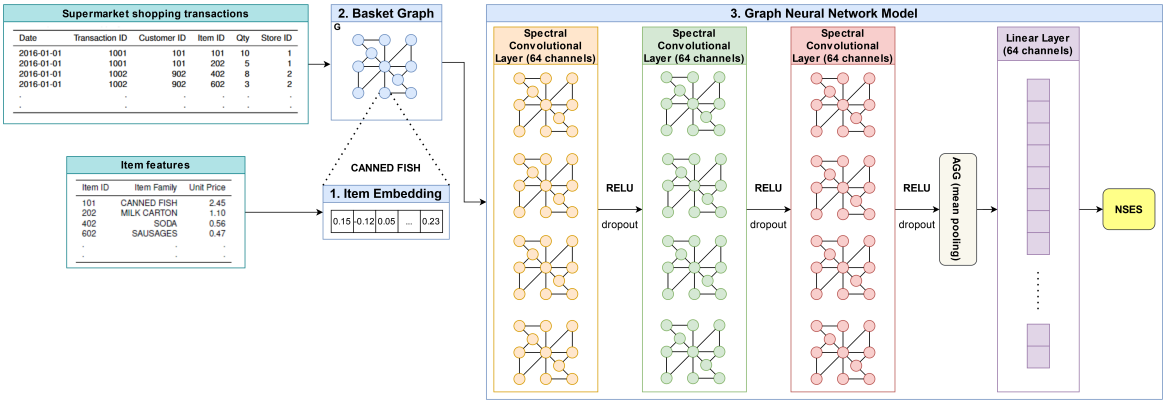


Fig. 4.5 The architecture of the proposed Graph Neural Network (GNN) model to predict the neighborhood socioeconomic status (NSES).

Chapter 5

Results

This chapter describes the performance indicators used to evaluate prediction power for the two machine learning models (a Regression Model and a Graph Neural Network Model). Additionally, it discusses the results and limitations of the proposed methodology to predict neighborhood socioeconomic status using top-up and supermarket transactions.

5.1 Model evaluation

5.1.1 Regression Model performance to predict NSES

The NSES is a unidimensional socioeconomic indicator, thus represented as a continuous value. In 2017, NSES values for urban neighborhoods in Guayaquil and Quito that had top-up transactions ranged from 128.37 USD to 755.62 USD. A logarithmic transformation was applied to the NSES values to enhance model performance. The effectiveness of the regression model was evaluated with two performance indicators: the coefficient of determination R^2, which measures the model's goodness-of-fit, and the root mean square error $RMSE$. During the training, the model was tested using different combinations of features (from 1 to 3) to evaluate how complexity affects performance. Results, summarized in Table 5.1, indicate that the model achieves optimal performance with two features: average denomination and denomination diversity. Linear Regression with regularization L2 (Ridge) has the best coefficient of determination and the lowest root mean square error. The best model was selected based on its simplicity and high performance, which can accurately predict the NSES (log) with a prediction rate of up to 74% for urban neighborhoods with an error of +/- 0.25. Figure 5.1 illustrates the Pearson correlation for the predicted and surveyed NSES, showing a linear correlation with a strong positive correlation $\rho = 0.85$ and a p-value of $1.83e-12$, which means the correlation is statistically significant. Therefore, the prediction

Table 5.1 Summary of models' performance.

Model	features	R^2	$RMSE$
Random Forest Regression	2	0.60	0.32
Linear Regression	2	0.71	0.27
Linear Regression with L1 (Lasso)	2	0.72	0.26
Linear Regression with L2 (Ridge)	2	0.74	0.25
Support Vector Regression	2	0.71	0.27

of the NSES can be represented as a linear combination of two consumption behavior indicators (average denomination and denomination diversity) (5.1).

$$\log(NSES) = \beta_0 + \beta_1(averageDenomination) + \beta_2(denominationDiversity) + e_i \quad (5.1)$$

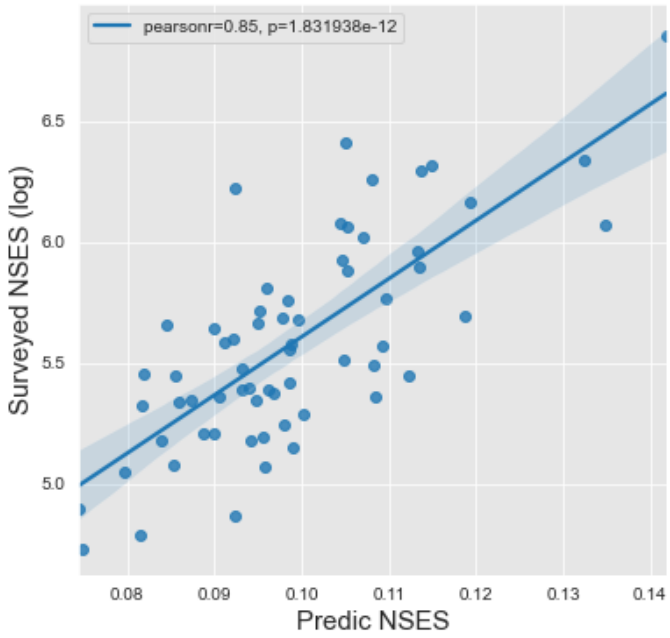


Fig. 5.1 Pearson correlation between predicted and surveyed NSES (log).

5.1.2 Graph Neural Network Model performance to predict NSES

GNN training and selection

The set of Basket Graphs samples from 2016 and 2017 were used to train and select the best model. The range of NSES values in Guayaquil's urban neighborhoods, based on supermarket transactions from 2016 to 2017, was between 138.39 USD and 575.87 USD. The GNN model's predictive power was evaluated using two performance indicators: R^2 and $RMSE$. Additionally, five-fold cross-validation was employed to assess the efficacy of both Chebyshev spectral convolutional filters (ChebConv) and various spatial convolutional filters (SAGEConv, GraphConv, GCNConv) to select the best model. The results, detailed in Table 5.2, show that the GNN model using the Chebyshev filter was the best model, achieving the highest $R^2 = 0.86$ and the lowest average $RMSE = 41.01$, indicating that the average error in the prediction of the NSES for the urban neighborhoods with different splits of training and validation samples is +/- 41.01 $.

Table 5.2 Results of five-fold cross-validation evaluation for spectral and spatial convolutional filters used by the GNN model.

	ChebConv		SAGEConv		GraphConv		GCNConv	
Fold #	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Fold 1	39.11	0.91	60.43	0.78	72.39	0.69	72.22	0.69
Fold 2	34.15	0.91	64.45	0.66	61.28	0.70	37.66	0.88
Fold 3	46.68	0.82	50.90	0.78	79.85	0.46	44.85	0.82
Fold 4	49.53	0.74	56.95	0.65	60.54	0.62	68.48	0.50
Fold 5	35.55	0.92	56.87	0.78	68.82	0.69	73.21	0.64
Avg	**41.01**	**0.86**	57.93	0.74	68.57	0.63	59.29	0.71
Std	6.09	0.07	4.48	0.06	7.21	0.09	14.98	0.13

GNN Evaluation

The model is evaluated using a set of basket graph samples from urban neighborhoods surveyed during the 2018 quarters and not used during model training and cross-validation. The range of NSES values in Guayaquil's urban neighborhoods, based on supermarket transactions from 2018, was between 134.76 USD and 600.35 USD. Table 5.3 details the errors between the model prediction and the surveyed value of the NSES for the Guayaquil urban neighborhoods that were surveyed during all quarters in 2018. The results reflect the highest error in the NSES prediction for the Bellavista neighborhood during all quarters, while Guayacanes and Garzota only in the first and second quarters. Bellavista and Garzota have

the highest NSES, and Guayacanes has a medium-high NSES. The error in prediction may be caused by the limited number of urban neighborhood samples with that socioeconomic level during the training. The GNN model can accurately predict the NSES with a prediction rate of up to 91% for the urban neighborhoods of Guayaquil during 2018 with an error of +/- 33.70 $. Table 5.4 details the GNN model performance results for each quarter of 2018. The results confirm that the GNN model is capable of predicting the NSES of the urban neighborhoods during quarters of a year that were not part of the model training.

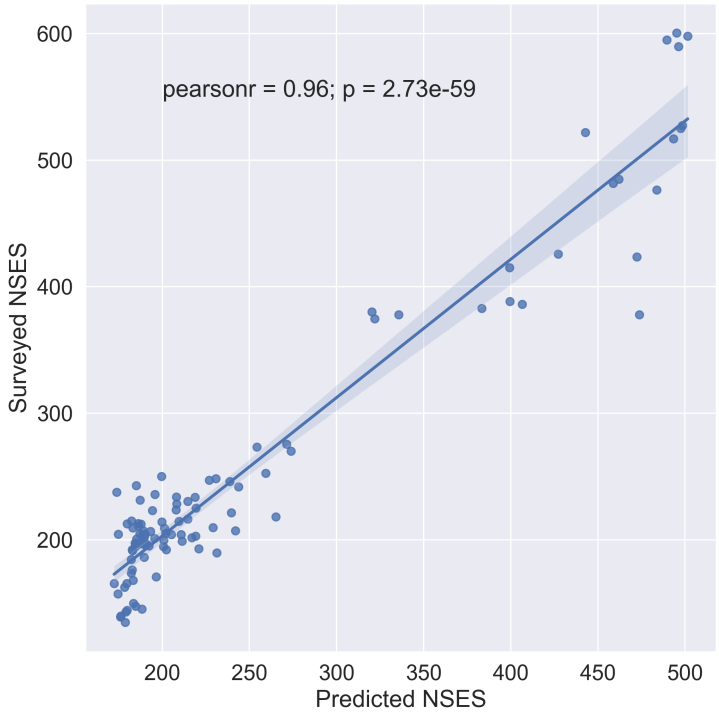


Fig. 5.2 Pearson correlation between predicted and surveyed NSES.

The correlation between the NSES predicted by the GNN model and the NSES surveyed by the INEC was computed for all urban neighborhoods in the test sample. Figure 5.2 illustrates this correlation, revealing a very strong positive correlation $\rho = 0.95$ with a statistically significant p-value of $6.82e-34$. This analysis confirms the capability of the GNN model using spectral convolutional filters to accurately predict surveyed NSES from the Basket Graphs for urban neighborhoods in Guayaquil during 2018.

The spatial distribution of predicted and surveyed NSES in Guayaquil during 2018 is shown in Figures 5.3. These choropleth maps categorize urban neighborhoods by socioeconomic levels and highlight the geographic trends observed during the quarters of 2018. The urban neighborhoods with the highest NSES are predominantly located in the

Table 5.3 Error measurement results between the predicted and surveyed NSES for Guayaquil urban neighborhoods that were surveyed during all quarters in 2018.

	Quarter 1				**Quarter 2**		
Urban neighborhood	NSES Surveyed	NSES Predicted	Error	Urban neighborhood	NSES Surveyed	NSES Predicted	Error
Bellavista	**589.71**	**496.22**	**93.49**	**Bellavista**	**594.88**	**489.55**	**105.33**
Garzota	516.74	493.42	23.32	**Garzota**	**521.91**	**442.81**	**79.10**
Samanes	476.48	483.86	-7.39	Samanes	481.65	458.75	22.90
Alborada	415.12	399.39	15.73	Guayacanes	382.98	383.39	-0.42
Guayacanes	**377.81**	**473.89**	**-96.08**	Sauces	374.66	321.87	52.78
Mucho Lote	241.91	243.83	-1.91	Parroquia Garcia Moreno	270.17	273.90	-3.73
Floresta	237.76	173.84	63.93	Mucho Lote	247.08	226.90	20.18
Mapasingue	225.22	219.30	5.92	Floresta	242.93	185.04	57.89
Febres Cordero	223.15	194.32	28.83	Mapasingue	230.39	214.58	15.80
Parroquia Letamendi	213.03	186.18	26.85	Febres Cordero	228.32	208.24	20.07
Guasmo Norte	204.42	174.63	29.79	Parroquia Letamendi	218.20	265.13	-46.93
Cristo del Consuelo	204.13	205.18	-1.05	Juan Montalvo	211.07	186.30	24.76
Abel Gilbert	198.97	211.28	-12.32	Guasmo Norte	209.59	183.01	26.58
Guasmo Sur	196.49	191.03	5.45	Cristo del Consuelo	209.30	201.15	8.16
Guasmo Oeste	196.13	189.70	6.43	Abel Gilbert	204.14	210.82	-6.68
Batallon del Suburbio	194.67	184.39	10.28	Guasmo Sur	201.66	216.88	-15.22
Vergeles	192.31	182.63	9.69	Guasmo Oeste	201.30	195.58	5.72
Puerto Lisa	192.25	202.32	-10.06	Batallon del Suburbio	199.84	200.86	-1.02
Pascuales	186.31	189.63	-3.32	Vergeles	197.48	184.21	13.28
Isla Trinitaria	184.59	182.06	2.53	Puerto Lisa	197.42	184.64	12.78
Bastion Popular	165.52	172.29	-6.76	Pascuales	191.48	182.75	8.73
El Fortin	157.29	174.40	-17.11	Isla Trinitaria	189.76	231.15	-41.39
Nueva Prosperina	139.17	175.98	-36.82	Bastion Popular	170.69	196.40	-25.71
Flor de Bastion	134.76	178.73	-43.97	El Fortin	162.46	178.23	-15.77
				Nueva Prosperina	144.34	179.74	-35.40
				Flor de Bastion	139.93	176.10	-36.17
	Quarter 3				**Quarter 4**		
Urban neighborhood	NSES Surveyed	NSES Predicted	Error	Urban neighborhood	NSES Surveyed	NSES Predicted	Error
Bellavista	**598.05**	**501.59**	**96.46**	**Bellavista**	**600.36**	**495.25**	**105.11**
Garzota	525.08	497.50	27.59	Garzota	527.39	498.47	28.92
Samanes	484.82	462.12	22.69	Alborada	425.77	427.26	-1.49
Alborada	423.46	472.30	-48.83	Guayacanes	388.46	399.44	-10.98
Guayacanes	386.15	406.53	-20.38	Sauces	380.14	320.34	59.79
Sauces	377.83	335.69	42.14	Parroquia Garcia Moreno	275.65	271.28	4.37
Parroquia Garcia Moreno	273.34	254.29	19.05	Mucho Lote	252.56	259.42	-6.85
Mucho Lote	250.25	199.58	50.67	Floresta	248.41	230.63	17.78
Floresta	246.10	238.75	7.35	Mapasingue	235.87	195.72	40.14
Mapasingue	233.56	218.77	14.79	Febres Cordero	233.80	208.01	25.79
Febres Cordero	231.49	187.14	44.35	Parroquia Letamendi	223.68	207.82	15.86
Parroquia Letamendi	221.37	239.51	-18.14	Juan Montalvo	216.55	214.59	1.95
Juan Montalvo	214.24	199.74	14.49	Guasmo Norte	215.07	182.27	32.79
Guasmo Norte	212.76	179.84	32.92	Cristo del Consuelo	214.78	209.53	5.25
Cristo del Consuelo	212.47	187.65	24.82	Abel Gilbert	209.62	229.02	-19.40
Abel Gilbert	207.31	241.90	-34.59	Guasmo Sur	207.14	188.94	18.20
Guasmo Sur	204.83	187.04	17.79	Guasmo Oeste	206.78	193.18	13.60
Guasmo Oeste	204.47	190.13	14.34	Batallon del Suburbio	205.32	201.90	3.42
Batallon del Suburbio	203.01	219.09	-16.08	Vergeles	202.96	189.65	13.31
Vergeles	200.65	185.19	15.46	Puerto Lisa	202.90	188.29	14.62
Puerto Lisa	200.59	189.38	11.22	Pascuales	196.96	187.32	9.64
Pascuales	194.65	200.52	-5.87	Isla Trinitaria	195.24	192.29	2.95
Isla Trinitaria	192.93	220.88	-27.95	Bastion Popular	176.17	182.62	-6.45
Bastion Popular	173.86	182.05	-8.19	El Fortin	167.94	183.16	-15.22
El Fortin	165.63	179.58	-13.94	Nueva Prosperina	149.82	183.45	-33.63
Nueva Prosperina	147.51	184.73	-37.22	Flor de Bastion	145.41	188.25	-42.84
Flor de Bastion	143.10	178.98	-35.88				

center and northeast, while the lowest NSES urban neighborhoods are found in the south and northwest.

Table 5.4 Summary of GNN models' performance.

Quarter	R^2	*RMSE*
1	0.91	35.09
2	0.89	36.47
3	0.92	32.57
4	0.92	30.53
All quarters	0.91	33.70

5.2 Discussion

The proposed methodology has both merits and demerits. For the merits, 1) The development of a Regression Model to capture the linear relationship between predictors of consumption behavior derived from the top-up transactions and the NSES; 2) The study proposed a graph representation to model consumption behavior using supermarket transactions, which can be used as an alternative data source to measure the economic rhythms of the population to overcome traditional data collection methods; 3) the design of a GNN model architecture, which learned the relationship between consumption behavior and the average per capita income could be used as a solution to estimate socioeconomic status at the neighborhood level. For the demerits, 1) it is necessary to re-train the Regression and GNN model when consumption patterns change significantly for disruptive events or urban neighborhood contexts different from those trained in the model. 2) it is difficult to interpret the learned graph embedding; therefore, it contrasts with the socioeconomic analysis, which usually requires a high interpretability of the models and results. Although the explanation of the predictions made by a GNN model remains an open problem, there are different studies that propose methodologies for the explanation of the prediction of a GNN model [40–42]. A GNN-Explainer model was trained to identify a compact subgraph structure that has a crucial role in predicting the NSES performed by the GNN model. Table 5.5 details the combinations of items that were purchased for Low and High NSES urban neighborhoods during all quarters of 2018. On the one hand, the items that stand out in the lowest NSES neighborhood are disposable tableware and cheap yogurts. These items highlight the economic reality of the neighborhood, where disposable tableware is commonly used for parties or fast food businesses, and cheap yogurts are used in breakfasts. On the other hand, in the urban neighborhood with the highest SES, different breakfast products, such as breakfast cereals and marshmallows, stand out. In addition, snacks are generally consumed between meals. Finally, dog foods stand out since customers buy processed pet foods in this type of urban neighborhood.

Table 5.5 The explanation of the prediction performed by the GNN model for Low and High NSES urban neighborhoods during all quarters of 2018

Quarter 1							
Low NSES				**High NSES**			
$Item_x$	$Price_x$	$Item_y$	$Price_y$	$Item_x$	$Price_x$	$Item_y$	$Price_y$
Toilet paper	0.70	Confectionery	0.34	Washing clothes	0.14	Powdered milk	0.62
Confectionery	0.34	**Yogurts**	0.18	Beef	0.54	**Breakfast cereals**	0.59
Desserts and sweets	0.65	Toilet paper	0.22	Whisky	0.19	Toilet paper	0.13
Flours	0.19	Biscuits	0.39	Bread	0.12	Juices and Tea	0.18
Desserts and sweets	0.56	Milk cream	0.10	Washing clothes	0.19	Fruits	0.34
Desserts and sweets	0.56	Spreadables	0.24	**Marshmallows**	0.15	Leaf vegetable	0.16
Flours	0.20	Hams	0.18	Juices and Tea	0.10	Chocolates	0.27
Juices and Tea	0.10	Toilet paper	0.22	Washing clothes	0.20	Biscuits	0.29
Liquid cleaners	0.35	Toilet paper	0.22	**Snacks**	0.90	Disposable tableware	0.76
Mortadella	0.15	Washing clothes	0.41	Whisky	0.19	Toilet paper	0.24
Quarter 2							
Low NSES				**High NSES**			
$Item_x$	$Price_x$	$Item_y$	$Price_y$	$Item_x$	$Price_x$	$Item_y$	$Price_y$
Sweeteners	0.30	Juices and Tea	0.20	Desserts and sweets	0.66	Fruits	0.37
Soft drinks	0.12	Fruits	0.21	Chicken	0.44	Condiments	0.21
Baby perfumes	0.42	Baby perfumes	0.31	Flavored drinks	0.20	Oatmeal	0.19
Hams	0.15	Oils	0.23	Chewing gum	0.43	Yogurts	0.14
Baby perfumes	0.45	Hair treatment	0.79	Liquid cleaners	0.22	Ice cream	0.22
Disposable tableware	0.60	Yogurts	0.50	Chewing gum	0.47	Hair dye	0.23
Beef	0.10	Ladyfingers	0.26	Desserts and sweets	0.66	Sweeteners	0.47
Shaving	0.27	**Yogurts**	0.15	Washing clothes	0.25	Soaps	0.28
Bathroom cleaners	0.23	Noodles	0.18	Hams	0.15	Condiments	0.13
Shampoo	0.82	Rice	0.14	Washing clothes	0.41	Fruits	0.32
Quarter 3							
Low NSES				**High NSES**			
$Item_x$	$Price_x$	$Item_y$	$Price_y$	$Item_x$	$Price_x$	$Item_y$	$Price_y$
Confectionery	0.39	Toilet paper	0.31	Confectionery	0.28	Rice	0.35
Disposable tableware	0.76	Toilet paper	0.31	Grains	0.10	Condiments	0.17
Pesticides	0.49	Carton milk	0.17	Mortadella	0.20	Desserts and sweets	0.56
Washing clothes	0.28	Yogurts	0.54	Mortadella	0.19	Biscuits	0.57
Flavored drinks	0.14	Whisky	0.23	Stains removal	0.37	Kitchen disposables	0.27
Desserts and sweets	0.56	Hair dye	0.23	Vodka	0.26	Rice	0.23
Chocolates	0.20	Flours	0.16	Noodles	0.45	Oils	0.16
Liquid cleaners	0.23	Hair dye	0.23	Toilet paper	0.70	Rice	0.15
Foot powder	0.42	Toilet paper	0.31	Mortadella	0.19	**Dog food**	0.10
Chocolates	0.20	**Yogurts**	0.10	Bagged dairy drinks	0.21	Juices and Tea	0.27
Quarter 4							
Low NSES				**High NSES**			
$Item_x$	$Price_x$	$Item_y$	$Price_y$	$Item_x$	$Price_x$	$Item_y$	$Price_y$
Whisky	0.17	Toothbrushes	0.21	Washing clothes	0.19	Air fresheners	0.14
Liquid cleaners	0.55	Fruits	0.42	Dried fruits	0.15	Soaps	0.19
Hams	0.10	Hair conditioner	0.58	Sanitary pads	0.29	**Marshmallows**	0.19
Washing clothes	0.28	**Disposable tableware**	0.60	**Marshmallows**	0.19	Coffee	0.39
Body deodorant	0.47	Margarines	0.53	Hams	0.19	**Breakfast cereals**	0.28
Candy	0.15	Chocolates	0.27	Baby perfumes	0.37	**Dog food**	0.23
Noodles	0.49	Juices and Tea	0.11	Bagged dairy drinks	0.21	Rice	0.11
Toilet paper	0.25	Toilet paper	0.19	**Snacks**	0.39	Canned foods	0.19
Sweeteners	0.47	Condiments	0.16	Dishwasher detergent	0.15	**Dog food**	0.23
Mortadella	0.15	Chocolates	0.20	Dried fruits	0.44	Rice	0.38

5.3 Limitations

Some limitations exist in utilizing top-up and supermarket transaction datasets to predict NSES. First, these datasets are not publicly available to estimate socioeconomic status. Nevertheless, the central and local governments of developing countries can reach agreements with the providers to access this anonymized data source to be used by our Regression and/or GNN model to measure the NSES with high accuracy and support their decision on how to allocate resources according to the needs of different geographical areas. In addition to the access, another limitation is the geographic distribution of the shops and supermarket stores across the city. The presence of shops and supermarket stores covers most of the urban neighborhoods; thus, measuring the customers' consumption behavior through our methodology for the urban neighborhoods where there are no shops and supermarket stores is not feasible. However, governments can take advantage of the agreements they can reach to access transactions with other providers in urban neighborhoods where they have a presence. Furthermore, regional limitations are present in our study because it was deployed to the urban neighborhoods of two cities in a developing country. One region's consumption patterns may differ from those of another region; therefore, evaluating the model's predictive power with top-up and supermarket transactions of other regions becomes essential.

Regarding the surveyed NSES, there is a limitation given by the urban neighborhoods that the INEC did not survey during the observation period. Consequently, NSES is unavailable for these urban neighborhoods, which can not be considered for the supervised learning process, even though the Regression and GNN model can infer NSES for those urban neighborhoods with shops and supermarket stores that can not be evaluated with the official NSES.

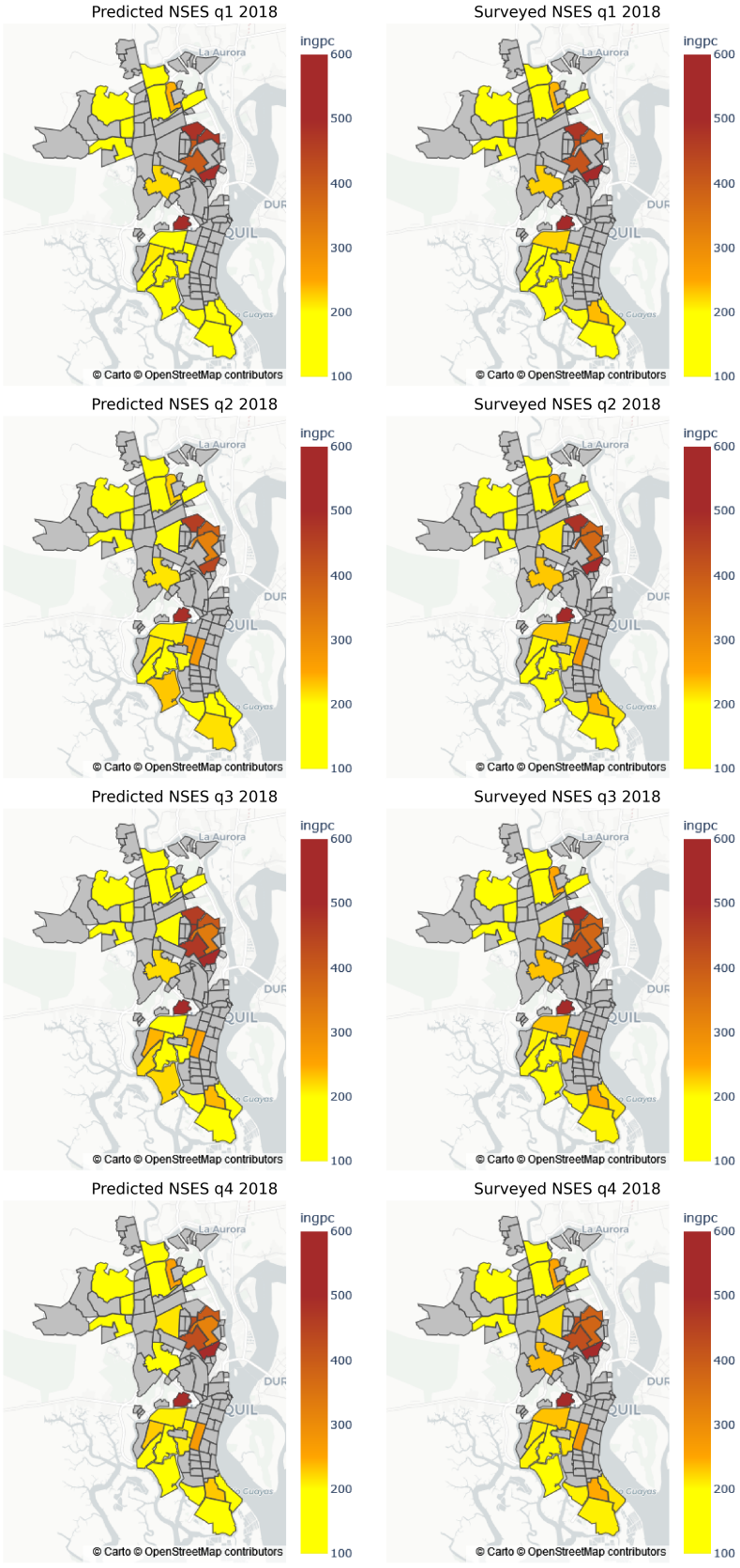


Fig. 5.3 Choropleth maps for the urban neighborhoods of Guayaquil used for testing the GNN model. The rows show the evaluation results for each quarter of 2018. The first column illustrates the predicted NSES values, and the second column Illustrates the surveyed NSES values.

Chapter 6

Conclusions and Future Work

This final chapter summarizes the significant contributions made in estimating socioeconomic status at the urban neighborhood level using digital sources to overcome traditional collection methods. Furthermore, future work is identified for the research community, which can explore new alternatives in measuring socioeconomic status at the intraurban level using digital sources.

6.1 Main findings

This dissertation proposes a Regression Model to predict socioeconomic status at the urban neighborhood level, using indicators of consumption behavior extracted from aggregated data of top-up transactions performed by customers in different shops located across urban neighborhoods of two populated cities in a developing country. The indicators of consumption behavior are good indicators of customer economic behavior and let us characterize the Neighborhood Socioeconomic Status; five machine learning algorithms were compared to model the relationship between top-up transactions and urban neighborhood wealth. The comparative analysis reveals that linear regression with regularization L2 (Ridge) is the best model to accurately estimate the NSES (log). The predictors related to the denomination, such as average denomination and denomination diversity, let us add significant predictive power to the model by predicting the socioeconomic status with a prediction rate of up to 74% for urban neighborhoods with an error of +/- 0.25.

Furthermore, this dissertation proposes a systematic process to predict Neighborhood Socioeconomic Status using supermarket transactions. The study characterizes customers' consumption behavior by creating a Basket Graph highlighting the more common combination of items purchased in an urban neighborhood to train a GNN model that learns the relationship between Basket Graph and NSES. The model's predictive power

was compared using spectral and spatial convolutional filers with cross-validation. The best result was the Chebyshev spectral convolutional filter, achieving the highest predictive power of $R^2 = 0.91$ and the lowest $RMSE = 33.70$ on average. This work reveals that characterizing the customers' consumption behavior from a graph perspective allows us to identify socioeconomic status at the urban neighborhood level. Moreover, the GNN model can accurately learn the relationship between the customers' consumption behavior and the NSES. The GNN model with spectral convolutional filters provides a new way to estimate urban neighborhood inequalities. Therefore, governments of developing countries can make informed decisions for resource allocation in poverty alleviation programs.

The findings of this study can be generalized to other developing countries with similar access to top-up and/or supermarket transactions. The methodology proposed can also be used as a low-cost tool for governments to make informed decisions on poverty resource allocation at the urban neighborhood level.

6.2 List of Publications

This thesis has led to the following publications:

- Cruz, E., Vaca, C., & Avendaño, A. (2019, December). Mining top-up transactions and online classified ads to predict urban neighborhoods socioeconomic status. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4055-4062). IEEE.

- Cruz, E., Vaca, C., & Villavicencio, M. (2021, December). Estimating urban socioeconomic inequalities through airtime top-up transactions data. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4265-4272). IEEE.

6.3 Future Work

Despite recent progress, there is still work to be done to enhance the estimation of socioeconomic status at the intra-urban level using digital sources. There are several potential research lines that could be pursued to improve the accuracy of predicting socioeconomic status at the urban neighborhood level. Some of the areas of improvement based on the methodology proposed in this dissertation are described below:

- To evaluate the performance of the Regression and GNN model with top-up and supermarket transactions from other regions to re-train the models with consumption patterns that could change significantly.

- To evaluate the performance of Multi-modal models in predicting poverty at the intra-urban level, combining the proposed Reggresion and GNN model with different machine learning and deep learning architectures using heterogeneous digital data sources to predict NSES with a higher predictive power than the presented in this study.

References

[1] UN. Ending poverty in all its forms everywhere. https://unstats.un.org/sdgs/report/2022/goal-01/, 2022. Accessed: 2024-02-10.

[2] Marco Hernandez, Lingzi Hong, Vanessa Frias-Martinez, Andrew Whitby, and Enrique Frias-Martinez. Estimating poverty using cell phone data: evidence from guatemala. *World Bank Policy Research Working Paper*, (7969), 2017.

[3] Jessica E Steele, Carla Pezzulo, Maximilian Albert, Christopher J Brooks, Elisabeth zu Erbach-Schoenberg, Siobhán B O'Connor, Pål R Sundsøy, Kenth Engø-Monsen, Kristine Nilsen, Bonita Graupe, et al. Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, 8(1):1–12, 2021.

[4] Xizhi Zhao, Bailang Yu, Yan Liu, Zuoqi Chen, Qiaoxuan Li, Congxiao Wang, and Jianping Wu. Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, 11(4):375, 2019.

[5] Guie Li, Zhongliang Cai, Yun Qian, and Fei Chen. Identifying urban poverty using high-resolution satellite imagery and machine learning approaches: Implications for housing inequality. *Land*, 10(6):648, 2021.

[6] Serena Giurgola, Simone Piaggesi, Márton Karsai, Yelena Mejova, André Panisson, and Michele Tizzoni. Mapping urban socioeconomic inequalities in developing countries through facebook advertising data. *arXiv preprint arXiv:2105.13774*, 2021.

[7] Jing Ma, Liangwei Yang, Qiong Feng, Weizhi Zhang, and Philip S Yu. Graph-based village level poverty identification. *arXiv preprint arXiv:2302.06862*, 2023.

[8] Rui Cao, Wei Tu, Jixuan Cai, Tianhong Zhao, Jie Xiao, Jinzhou Cao, Qili Gao, and Hanjing Su. Machine learning-based economic development mapping from multi-source open geospatial data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:259–266, 2022.

[9] Florina Pînzaru, Adina Săniuță, and Bianca R Sălăgeanu. Managing innovation for sustainability in public administration: the challenges of capacity-building. *Nowoczesne Systemy Zarządzania*, 17(3):65–79, 2022.

[10] Christoph Lakner, Daniel Gerszon Mahler, Mario Negre, and Espen Beer Prydz. How much does reducing inequality matter for global poverty? *The Journal of Economic Inequality*, 20(3):559–585, 2022.

[11] WorldBank. Poverty overview. https://www.worldbank.org/en/topic/poverty/overview, 2023. Accessed: 2024-02-10.

[12] ECLAC. Extreme poverty in the region rises to 86 million in 2021 due to the deepening of the social and health crisis prompted by the covid-19 pandemic. https://www.cepal.org/en/pressreleases/extreme-poverty-region-rises-86-million-2021-due-deepening-social-and-health-crisis, 2022. Accessed: 2024-02-10.

[13] Surendran Padmaja Subash, Rajeev Ranjan Kumar, and Korekallu Srinivasa Aditya. Satellite data and machine learning tools for predicting poverty in rural india. *Agricultural economics research review*, 31(2):231–240, 2018.

[14] Guberney Muñetón-Santa and Luis Carlos Manrique-Ruiz. Predicting multidimensional poverty with machine learning algorithms: an open data source approach using spatial data. *Social Sciences*, 12(5):296, 2023.

[15] Eduardo Cruz, Carmen Vaca, and Allan Avendaño. Mining top-up transactions and online classified ads to predict urban neighborhoods socioeconomic status. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4055–4062. IEEE, 2019.

[16] Eduardo Cruz, Carmen Vaca, and Monica Villavicencio. Estimating urban socioeconomic inequalities through airtime top-up transactions data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4265–4272. IEEE, 2021.

[17] Ysabel Atiencia, Eduardo Cruz, Carmen Vaca, and Lessette Zambrano. Spatio-temporal analysis: Using instagram posts to characterize urban point-of-interest. In *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 114–119. IEEE, 2020.

[18] Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, and Piero Fraternali. Taking brazil's pulse: tracking growing urban economies from online attention. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 451–456. ACM, 2014.

[19] Xiaowen Dong, Alfredo J Morales, Eaman Jahani, Esteban Moro, Bruno Lepri, Burcin Bozkaya, Carlos Sarraute, Yaneer Bar-Yam, and Alex Pentland. Segregated interactions in urban and online space. *EPJ Data Science*, 9(1):20, 2020.

[20] José Carpio-Pinedo, Gustavo Romanillos, Daniel Aparicio, María Soledad Hernández Martín-Caro, Juan Carlos García-Palomares, and Javier Gutiérrez. Towards a new urban geography of expenditure: Using bank card transactions data to analyze multi-sector spatiotemporal distributions. *Cities*, 131:103894, 2022.

[21] Yongming Xu, Yaping Mo, and Shanyou Zhu. Poverty mapping in the dian-gui-qian contiguous extremely poor area of southwest china based on multi-source geospatial data. *Sustainability*, 13(16):8717, 2021.

[22] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[23] Office for National Statistics. Economic activity and social change in the uk, real-time indicators: 15 june 2023. https://www.ons.gov.uk/economy/economicoutputandproductivity/output/bulletins/economicactivityandsocialchangeintheukrealtimeindicators/15june2023, 2023. Accessed: 2024-02-10.

[24] UK FINANCE. Uk payment markets summary 2023. https://www.ukfinance.org.uk/system/files/2023-09/UK%20Finance%20Payment%20Markets%20Report%202023%20Summary.pdf, 2023. Accessed: 2024-02-10.

[25] Banco Central del Ecuador. Preferencia de medios de pagos por parte de los hogares ecuatorianos. https://contenido.bce.fin.ec/documentos/Administracion/boletinInfSNPV01.pdf, 2019. Accessed: 2024-02-10.

[26] ARCOTEL. ParticipaciÓn de mercado e Índice de concentraciÓn de los servicios de telecomunicaciones. https://www.arcotel.gob.ec/wp-content/uploads/2021/12/Boleti%CC%81n-estadistico-noviembre-2021.pdf, 2021. Accessed: 2024-02-10.

[27] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690, 2017.

[28] Peter Yamakawa, Gareth H Rees, José Manuel Salas, and Nikolai Alva. The diffusion of mobile telephones: An empirical analysis for peru. *Telecommunications Policy*, 37(6-7):594–606, 2013.

[29] Patrick Hosein, Gabriela Sewdhan, and Aviel Jailal. Soft-churn: Optimal switching between prepaid data subscriptions on e-sim support smartphones. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–6. IEEE, 2021.

[30] NU CEPAL. State of broadband in latin america and the caribbean 2017. 2018.

[31] INEC. Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) Documento Metodológico. https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Septiembre-2018/ENEMDU_Metodologia%20Encuesta%20Nacional%20de%20Empleo%20Desempleo%20y%20Subempleo.pdf, 2018. Accessed: 2024-02-10.

[32] Pål Sundsøy, Johannes Bjelland, Bjørn-Atle Reme, Asif M Iqbal, and Eaman Jahani. Deep learning applied to mobile phone data for individual income classification. In *2016 International Conference on Artificial Intelligence: Technologies and Applications*. Atlantis Press, 2016.

[33] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[34] Luca Pappalardo, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. Using big data to study the link between human mobility and socio-economic development. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 871–878. IEEE, 2015.

[35] Yang Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems*, 2018.

[36] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[37] Ivan F Videla-Cavieres and Sebastián A Ríos. Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, 41(4):1928–1936, 2014.

[38] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[39] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[40] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33:19620–19631, 2020.

[41] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[42] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for nlp with differentiable edge masking. *arXiv preprint arXiv:2010.00577*, 2020.