Journal of Information Systems Engineering and Management
2025, 10(20s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid Model of Machine Learning with Web Scrapping to Determine the Urban Dynamics of the Constructions of Medellín - Colombia

J.A. Castillo*[1], J.P. Barrero[2], Y. Ceballos[3]

[1]*Ingeniería y Sociedad Research Group, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia*
[2]*Universidad de Antioquia, Medellín, Colombia*
[3]*Ingeniería y Tecnologías de las Organizaciones y de la Sociedad (ITOS) Research Group, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia*

ARTICLE INFO

Received: 15 Dec 2024

Revised: 29 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

The cities growth has its main reason that more than 50% of the world's population lives in urban areas and in Latin America that percentage exceeds 75%. In many aspects, this growth is studied as a territory-oriented manner, increasing the area of occupation only in two dimensions; cities grow in territory and volume. The present research focuses on vertical growth from a view composed of two independent models, the first uses a classification Machine Learning model using statistical values of the last ten years of changes in story levels and other variables to predict if a property would present a vertical growth, the other model is created with information taken from the extraction using Web Scrapping technics from the market offers from a popular real state web page in Colombia. These two data are projected geographically using the neighborhoods of Medellín and through a raster model of average density, the changes are added to provide a result by neighborhoods of the city of Medellín where there will be the highest probability of new stories or new buildings, the goal is to improve the management of the territory in terms of contributing to the study of the city and the fiscal and cadastral aspects of the territory.

Keywords: Machine Learning; Arcgis; Arcpy; Vertical Growth; Web Scrapping; Logistic Regression.

INTRODUCTION

The cities growth has been highly studied and the principal reason is that more than 50% of the world's population lives in urban areas and in Latin America that percentage exceeds 75% [1]–[3]. In many aspects, the growth of cities is studied in a view of territory, increasing the area of occupation only in two dimensions [4]. Cities change and grow mainly in two areas, in territory and in volume, construction is a fundamental part of the dynamism of a city providing housing and work options by increasing their size vertically [5], [6]. The current model of prediction of vertical growth of the city of Medellín is based on linear statistical projections of occupation and metrics of perception in territory through field visits by government entities.

Therefore, it is pertinent to identify the sectors where the supply and demand of new constructions is higher than the other locations in the urban area, the objective is to create a hybrid model of machine learning[7], [8] and geographic information systems such as ArcGIS and QGIS [9], [10] where it allows to identify the sectors of the city where vertical growth has greater values compared to historical physical changes and real estate offers of the urban area, finally we combine the two layers and the neighborhoods with the greatest influence or probability of vertical growth.

Urban dynamics play a vital role in shaping the growth and development of cities worldwide. In recent years, the availability of large datasets and the emergence of powerful Machine Learning (ML) algorithms have led to significant advances in urban studies. This paper presents a hybrid model of ML with web scraping to determine the urban dynamics of the constructions of Medellín, Colombia.

The city of Medellín, located in the Aburrá Valley, has undergone significant changes in its urban landscape over the past few decades. With a population of over 2.5 million, the city has experienced rapid urbanization, which has resulted in the construction of numerous buildings, roads, and other urban infrastructure.

This study aims to analyze the urban dynamics of construction in Medellín using a hybrid model of ML with web scraping. The model will be used to gather data from various sources, including official government websites and social media platforms, to analyze and predict the patterns of urban construction in the city.

The paper will present the methodology used to collect and analyze the data, including the ML algorithms employed. The results of the study will be presented, highlighting the significant trends and patterns in the urban dynamics of construction in Medellín. Finally, the paper will conclude with a discussion of the implications of the findings and the potential applications of the hybrid model of ML with web scraping in urban studies.

THEORETICAL FRAMEWORK

The model requires two types of data from different sources, which are obtained from government information and the other obtained from the web. The data is shown as follows: Information on physical changes and predominant social class of the urban area [11] and the later are the real estate offers taken through Web Scrapping (WS) of the website FincaRaiz.com using Selenium [12], [13]. Cadastral information is a fundamental for the present study, because it has the data to develop the model using the field sizes as the level of aggregation; it is known also as the property unit, where the lot or soil is the unit on which it is intended to identify a construction in it and which could change over time with a new construction or building. The source of information are the alphanumeric and historical geographical cadastral databases from 2011 to 2020, provided by the "*Subsecretaría de Catastro*", which for this study requests compliance with the guidelines established by Law 1266 of 2008 [14], which defines the protection of the personal data of owners and holders. The cadastral database includes the information of owners, property, addresses and physical information of floor levels (stories) which is used in this model. The cadastral information is shown in Table 1 for a total of 169356 records from the years 2011 to 2020.

Table 1. Cadastral Information

Field	Detail
Comuna	Comuna refers to an administrative unit that groups specific sectors or neighborhoods
CBML	It refers to the classification code for a terrain in a specific neighborhood.
Mode Economic Use	More repeated economic use in a CBML. Economic use is defined as the type of ownership, e.g. housing, commerce, industry, etc.
Description Use	Text describing mode use
Available	Boolean variable that details True if the terrain does not have constructions and False if it has them
Terrain Area	Value in m^2 of the terrain area
Construction Area	Value in m^2 of the construction area
Increase	Boolean value detailing whether the property grew vertically in the reported years
Enable	Boolean value detailing whether the row meets the completeness requirements
Increase In the last period	Boolean value that details if the property grows in the last statistical record
Registry type	Record type characteristic
Value	Number of floors in a property
Year	Year of registration

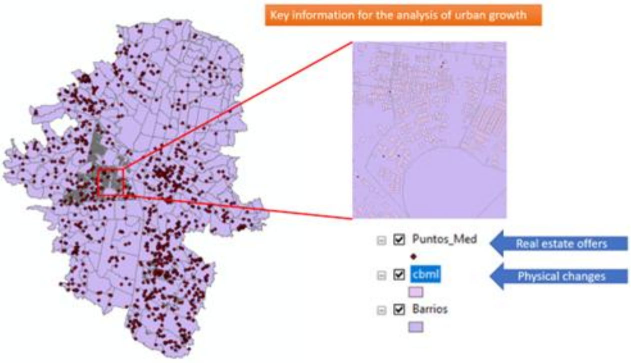


Figure 1. Detailed information about the real estate offers and the physical changes within the city

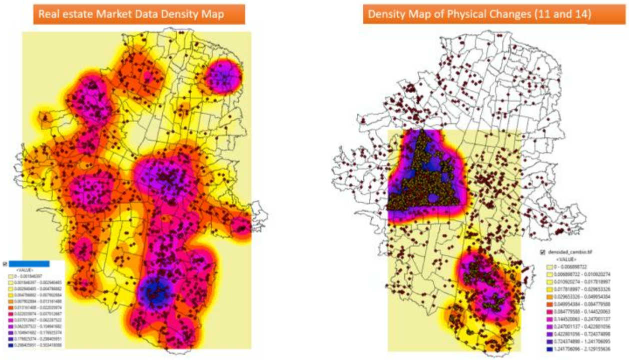


Figure 2. Real estate market data density map and density map of physical changes (ID locations 11 and 14)

The previous information must be complemented with the data from the real estate market no longer than ten years to be consistent with the data of physical changes of the cadastral information. This data was extracted using the Python library Selenium (Python.org, 2021; Thoughtworks, 2022) using a search with the keywords "Medellín - Antioquia" on the website Fincaraiz.com (Fincaraiz, 2022) obtaining 1820 offers of the real estate market in the urban area of Medellín. A validation process was made on the data to ensure the quality of the later in the process, after this process a total of 1584 offers are preserved. One key validation process was to check using ArcGIS the correctness of the "X" and "Y" coordinates and validate that the real state offers are in Medellín.

METHODOLOGY

The present model is developed in three phases, the first one consists of a Machine Learning (ML) model to determine the locations in the city where the probability of having a vertical growth is higher, the second consists of performing the geographical processing of the data from the ML model and the data from the real estate offers obtained from the WS and with georeferenced values to finish with a raster calculator of these two layers to identify the neighborhoods or locations with a higher probability of vertical growth. As shown in Fig 1, the points of both the offers and the physical changes are georeferenced to detail the behavior and the relationship between them. There is a high densification of points in the southeast sectors corresponding to the location called as "El Poblado" with the ID code 14 and the central western sector from where the zoom corresponding to ID code 11 called as "Laureles" is shown, using these points on the map we create the raster model to calculate the locations with higher grow probabilities.

The third phase of the model involves the use of web scraping techniques to gather additional data on the identified neighborhoods or locations. This includes data on population density, infrastructure, land use, and other relevant factors that may affect the urban dynamics of construction. The data collected is then processed using ML algorithms to generate predictions on future trends in urban development in these areas. The use of web scraping and ML algorithms in this model allows for a more comprehensive analysis of the urban dynamics of construction in Medellín. By combining data from various sources and using advanced analytical techniques, the model can provide insights into the patterns and trends of urban development that would be difficult to obtain through traditional methods.

One of the main advantages of this hybrid model is its ability to provide real-time data on urban development in Medellín. By continuously monitoring and analyzing data from various sources, the model can generate up-to-date predictions on the future growth and development of the city. This can be extremely useful for urban planners, policymakers, and other stakeholders in making informed decisions on urban development strategies.

Furthermore, the results of this study can have broader implications for urban studies in general. The use of web scraping and ML algorithms to analyze urban dynamics can be applied to other cities and regions, providing valuable insights into the complex interactions between various factors that shape urban development. As such, this study represents a significant contribution to the growing body of research on urban studies and highlights the potential of advanced analytical techniques in addressing complex urban challenges.

Machine Learning Model

We used Logistic Regression Classification (LRC), this model is used after testing multiple classification models, LRC provides the highest degree of accuracy in the prediction. one of the main problems in classification occurs when the algorithm never converges in updating the weights while being trained. This occurs when classes are not perfectly linearly separable. Therefore, to deal with binary classification problems, LRC is one of the most used algorithms and is a simple, but powerful classification algorithm (despite its name). It works very well in linearly separable classes and can be extended to multiclass classification [15]–[17]. For the present case, the changes in the values from ID locations 11 (Laureles) and 14 (Poblado) are taken, this because the points with greater relevance in vertical growth were detailed for these locations. The quality of the model is identified with the Method of Cross Validation with k = 3, This consists of dividing the data randomly into k groups of the same size, k-1 groups are used to train the model and one of the groups is used for validation. This process is repeated k times using a different group as validation in each iteration. For the current model, the validation scores were estimated around 0.8 showing good performance and quality of the model used. All these data are reflected in Fig 2 where the heat map shows the relationship between the territory and the result values of the ML model.

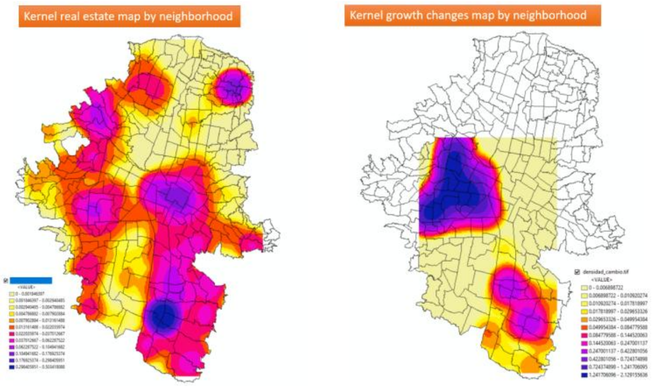


Figure 3. Kernel real estate map by neighborhood and Kernel growth changes map by neighborhood

Web Scrapping

Web scraping is a valuable tool in the field of data science and is widely used to gather data from various online sources. In the context of this study, web scraping was used to collect data from real estate websites in Medellín, providing a comprehensive view of the real estate market in the city. By scraping data from multiple websites, the model was able to generate a more accurate and up-to-date picture of the market, which was then used to inform predictions about urban development.

One of the key advantages of web scraping in this context is the ability to gather data that would be difficult or time-consuming to collect through other methods. Real estate websites provide a wealth of information on property prices, locations, and other features that can be used to identify patterns and trends in the market. By automating the process of data collection, web scraping allows for large amounts of data to be gathered quickly and efficiently, providing a more comprehensive view of the market.

In addition, web scraping can help to overcome limitations in the data provided by other sources. For example, the ML model used in this study relied on historical data on the behavior of properties at the terrain level to make predictions about future growth patterns. However, this data may not always be up-to-date or accurate. By supplementing this data with information from the real estate market gathered through web scraping, the model can generate more accurate predictions and identify any discrepancies between the model and the real estate market.

Overall, the use of web scraping in this study demonstrates the value of this technique in generating insights and informing decision-making in the field of urban development. By providing a more comprehensive and accurate view of the real estate market in Medellín, web scraping contributes to the accuracy and reliability of the hybrid model of machine learning used in this study to determine the urban dynamics of construction in the city.

Web Scrapping (Caballero et al., 2018) is the technic used to extract data from websites. Web scraping software can directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a user, the term usually refers to automated processes implemented using a robot or web crawler. In this way, data is collected in a spreadsheet to process and obtain high-quality information related to real estate date quickly and efficiently. The scraping is done to check if the sale of real estate in the same periods of time corresponds to the creation of new properties or to vertical growth. This validation methodology corresponds to the validation used in simulation where reality determines the behavior of the model [18]. The data obtained are as follows: City, Property Name, Neighborhood, State, Price, Rooms, Bathrooms, Parking, Built Area, Private Area, Social class level, Construction State, Antiquity, Floor Number (stories number), Administration Value, Price per m^2, XY (location), Other Features, URL and Date of execution. From the previous data will be used the values with less than 10 years in antiquity that are related according to the values of physical changes, this information can be visualized in the Fig 3, the process was made using Python and the Selenium library [13], [19].

Geographic Processing Model

For the geographic process we start with the assignment of the data frames necessary for the processing of the Shapefile type data. The layers to be used were obtained from the website of the Geographical Catalog of Medellín [20] where the layers of terrains and Neighborhoods are downloaded to contrast the information of the ML model and the WS data. The procedure to be performed for the parameters detailed above is as follows:

- WS data with XY value (Latitude and Longitude) is converted to geographic points.
- The properties resulting from the ML model are converted into points according to the centroid of the terrain where they are located.
- A density map (Kernel) is made for the real estate points.
- A density map (Kernel) is made for the points obtained from the ML model.
- A kernel cut is made for the real estate points and for the points of the ML model based in the neighborhoods through two independent processes.
- The average density of market points per neighborhood is calculated for the market values and data resulting from the ML model.

The algorithm is implemented using Esri's ArcPy library [9], [21]. The graphics resulting from the procedure are shown below in Fig 3. The final process is to make the sum of the two raster layers previously calculated to determine the neighborhoods of ID Locations 11 and 14 where greater growth is expected in terms of new constructions as shown in Fig 4, Fig 5 and Fig 6.

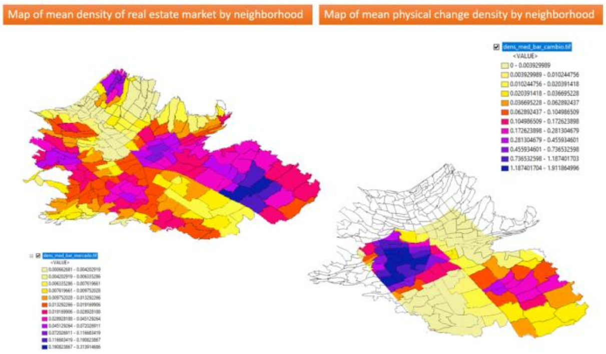
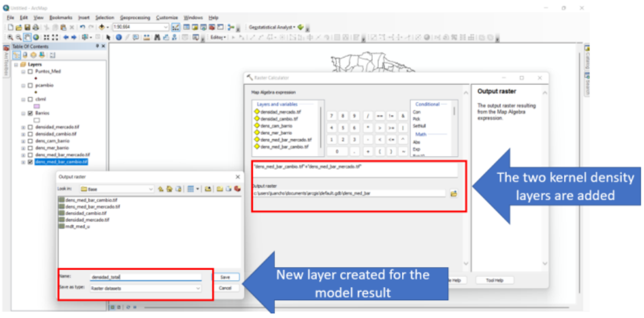


Figure 4. Map of mean density of real estate market by neighborhood and map of mean physical change density by neighborhood with 3D view.

Raster calculator for the kernel density map of urban growth



"dens_med_bar_cambio.tif"+"dens_med_bar_mercado.tif"

Output raster: c:\users\juancho\documents\arcgis\default.gdb\dens_med_bar

Name: densidad_total

Save as type: Raster datasets

The two kernel density layers are added

New layer created for the model result

Figure 5. Raster calculator for model results

Result of the raster sum map and kernel density

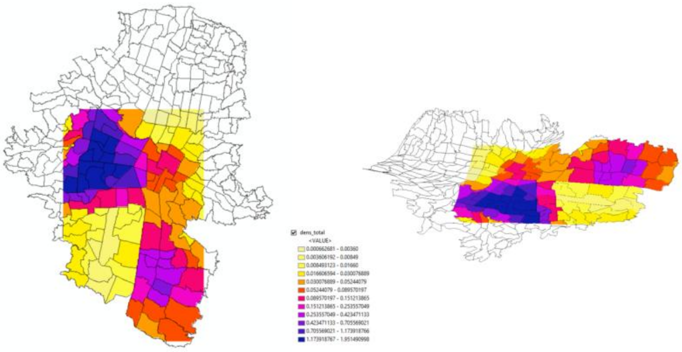


Figure 6. Results of the model for the ID Location 11 and 14 with 3D view

CONCLUSION

The composition of Fig 4, Fig 5 and Fig 6 shows the procedure to obtain the results where the neighborhoods are most likely to present a growth in the number of properties constructions using a logistic regression ML model, the capture of real estate offers as a method of cross-validation for their subsequent combination for a raster adding calculated density layers in Fig 5 and Fig 6 where the results show us several clusters with a higher probability of obtaining new constructions. These clusters are formed in areas with a dark blue color where for the sector of the ID Location for Laureles (11) is made up of the neighborhoods named "La Castellana", "Las Acacias", "Lorena," "Bolivariana", "San Joaquín", "La América", "El Velódromo" and "Florida Nueva". While in "El Poblado" the cluster is formed by the neighborhoods "Patio Bonito", "La Florida", "Alejandría", "Los Balsos Numero 2" and "La Aguacatala"; these results will be used by local administration to identify and to perform the cadastral visits on field.

The ML model in which the vertical growth values of the city of Medellín are determined through the historical behavior of the properties at the terrain level of the city has made it possible to identify sectors on which the property management units of the municipality of Medellín can provide an accompaniment to the cadastral processes which will contribute to the validation of the data using a comparison between the offer of the real estate market and the results of the locations in which in the same period of time real estate offers have been made providing a second layer of confidence to the model since we cannot trust only the ML model without having a direct comparison against the real estate market that allows to improve the forecast in a positive feedback.

The ML model used in this study provides a unique approach to identifying sectors of the city with high potential for vertical growth. By analyzing the historical behavior of properties at the terrain level, the model can generate accurate predictions on future patterns of urban development. This can be extremely useful for property management units in the municipality of Medellín, as it allows them to provide accompaniment to cadastral processes in sectors with higher growth potential. This can help to ensure that the appropriate regulations and policies are in place to manage urban development effectively.

In addition to identifying potential growth areas, the model also provides a means of validating the data through a comparison with the real estate market. By comparing the results of the ML model with the offers made in the same period, a second layer of confidence is added to the model. This is crucial since it allows for the improvement of the forecast through positive feedback. This validation process can also help to identify any discrepancies between the model and the real estate market, providing an opportunity to refine the model further.

Overall, the use of ML algorithms and web scraping techniques in this study represents a significant advance in the field of urban studies. By providing a more comprehensive and accurate analysis of the urban dynamics of construction in Medellín, this model can help to inform urban development policies and strategies. It also highlights the potential of advanced analytical techniques in addressing complex urban challenges and provides a framework for future research in this field.

REFERENCES

[1] P. B. R. Campos, C. M. de Almeida, and A. P. de Queiroz, “Educational infrastructure and its impact on urban land use change in a peri-urban area: a cellular-automata based approach,” *Land use policy*, vol. 79, no. August, pp. 774–788, 2018, doi: 10.1016/j.landusepol.2018.08.036.

[2] W. Cao, Z. Zhu, Y. Zhou, L. Liang, X. Li, and B. Yu, “Mapping annual urban dynamics (1985–2015) using time series of Landsat data,” *Remote Sens. Environ.*, vol. 216, no. August, pp. 674–683, 2018, doi: 10.1016/j.rse.2018.07.030.

[3] H. Dadashpoor, P. Azizi, and M. Moghadasi, “Analyzing spatial patterns, driving forces and predicting future growth scenarios for supporting sustainable urban growth: Evidence from Tabriz metropolitan area, Iran,” *Sustain. Cities Soc.*, vol. 47, no. March, p. 101502, 2019, doi: 10.1016/j.scs.2019.101502.

[4] Y. Fu *et al.*, “Characterizing the spatial pattern of annual urban growth using time series Landsat imagery,” *Sci. Total Environ.*, vol. 666, pp. 274–284, 2019, doi: 10.1016/J.SCITOTENV.2019.02.178.

[5] F. S. K. Agyemang and E. Silva, “Simulating the urban growth of a predominantly informal Ghanaian city-region with a cellular automata model: Implications for urban planning and policy,” *Appl. Geogr.*, vol. 105, no. August 2017, pp. 15–24, 2019, doi: 10.1016/j.apgeog.2019.02.011.

[6] J. He, C. Li, J. Huang, D. Liu, and Y. Yu, “Modeling Urban Spatial Expansion Considering Population Migration Interaction in Ezhou, Central China,” *J. Urban Plan. Dev.*, vol. 145, no. 2, 2019, doi: 10.1061/(ASCE)UP.1943-5444.0000503.

[7] S. C. Crommelinck and M. N. Koeva, “Towards Cadastral Intelligence?: Extracting visible boundaries from UAV data through image analysis and machine learning,” *GIM Int.*, 2019.

[8] V. K. Vemuri, “The Hundred-Page Machine Learning Book.” Taylor \& Francis, 2020.

[9] Esri Inc., “ArcGIS Desktop: Release 10.8.1,” *Redlands CA*. ESRI, Redlands, California, 2022. [Online]. Available: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

[10] QGIS Development Team, “QGIS.” Open Source Geospatial Foundation Project, 2022. [Online]. Available: https://qgis.org/en/site/.

[11] Alcaldía de Medellín, “Estadísticas catastrales,” *2022*, 2022, [Online]. Available: https://www.medellin.gov.co/irj/portal/medellin?NavigationTarget=contenido/4918-Estadisticas-catastrales-de-Medellin

[12] Fincaraiz, “Fincaraiz,” *2022*, 2022. https://fincaraiz.com.co/ (accessed Jun. 16, 2022).

[13] Thoughtworks, “Selenium.” Thoughtworks Holding, Inc., Chicago, Illinois, U.S., 2022. [Online]. Available: https://www.selenium.dev/about/

[14] Senado de la Republica de Colombia, “Ley 1266 de 2008.” Senado de la Republica de Colombia, Bogota, 2008.

[15] V. Roman, “Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos.” 2021. Accessed: Sep. 20, 2021. [Online]. Available: https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducción-a-la-clasificación-y-principales-algoritmos-dadee99c9407

[16] S. Raschka and V. Mirjalili, “Python Machine Learning: Machine Learning and Deep Learning with Python,” *Scikit-Learn, TensorFlow. Second Ed. ed*, 2017.

[17] J. Bobadilla, *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*. Ediciones de la U, 2021.

[18] J. Banks, *Discrete Event system Simulation*. 2010. doi: 10.1016/b0-12-227240-4/00045-9.

[19] Python.org, “Python 3.9,” 2021. https://docs.python.org/3.9/contents.html (accessed Sep. 02, 2021).

[20] Alcaldía de Medellín, “Catalogo Geográfico de Medellín,” *2022*, 2022. https://www.medellin.gov.co/giscatalogacion/srv/spa/catalog.search#/home (accessed Jun. 26, 2022).

[21] E. Christian Harder, *The ArcGIS Book*, Second Edi. 2019. [Online]. Available:

https://downloads.esri.com/LearnArcGIS/pdf/instructional-guide-for-the-arcgis-book-2e.pdf