Policy Research Working Paper 10301

Estimating House Prices in Emerging Markets and Developing Economies

A Big Data Approach

Daniela M. Behr
Lixue Chen
Ankita Goel
Khondoker Tanveer Haider
Sandeep Singh
Asad Zaman

WORLD BANK GROUP
International Finance Corporation
February 2023

Policy Research Working Paper 10301

Abstract

Despite the relevance of house prices for a variety of stakeholders as well as for macroeconomic and monetary policy making, reliable, publicly available house price data are largely absent in emerging markets and developing economies. Filling this void, this paper presents a systematic approach to collecting, analyzing, and assessing private property prices in emerging markets and developing economies. The paper uses data scraped from five countries' largest real estate websites where private properties are listed for sale, to obtain price data and property attributes to establish a comprehensive data set that allows for both intra- and inter-country comparison of residential property prices. It then outlines the usability of these data by employing random forest estimation to predict the price of a standard housing unit—the basic house price—that is comparable across countries. While this approach is also applicable to filling wide data gaps in the provision of private property prices in developed economies, the paper focuses on how this approach can be applied to emerging markets and developing economies, where private property price data are particularly scarce.

This paper is a product of the Development Impact Measurement Department, International Finance Corporation. It is part of a larger effort by the World Bank Group to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dbehr@worldbank.org; khaider@ifc.org; ssingh13@ifc.org; and azaman@ifc.org.

Produced by the Research Support Team

Estimating House Prices in Emerging Markets and Developing Economies: A Big Data Approach

Behr, Daniela M.; Chen, Lixue; Goel, Ankita; Haider, Khondoker Tanveer; Singh, Sandeep; Zaman, Asad*[1]

Keywords: **house price, web scraping, random forest, machine learning, residential real estate**

JEL classification: **R31 E30 C80**

[1] The paper was written while the authors were at the International Finance Corporation (IFC). This paper is part of a larger research project on affordable housing at IFC's Development Impact Measurement Department. Sandeep Singh, Khondoker Haider and Asad Zaman are Economists at IFC. Daniela Behr is currently an Economist at the World Bank. Lixue Chen and Ankita Goel are currently Research Assistants at the IMF. *Corresponding author: ssingh13@ifc.org.

The paper has benefitted from thoughtful guidance and encouragement by Issa Faye, Dan Goldblum, Camilo Mondragon-Velez, Zekebweliwai Geh, and participants of the IFC Sector Economics research sounding board. We want to thank our peer reviewers Adrian Alter, Nina Biljanovska, Jean-Charles Bricongne, Rui Costa, Niall O'Hanlon, Jane Jenkinson, Dunstan Matekenya, Nisachol Mekharat, Maurice Nsabimana, Edie Purdie, Marko Olavi Rissanen, Inyoung Song, Nouhoum Traore, and Simon Walley. We also received very helpful feedback and comments throughout the process from Jeffrey David Anderson, Richard Chamboko, Friedemann Roy, Ronald Kai Him Leung, Justice Mensah, and Philip Schellekens. Views expressed here are those of the authors, and all errors remain our own.

1. Introduction

Private property price levels and their movement over time have critical implications for the economy of most countries, but they also play a fundamental role in household (HH) wealth. Buying property is undoubtedly the most significant investment for many families around the world and reliable information on private property prices and their determinants are essential decision-making factors for HHs. The relationship between property prices and HHs' disposable incomes determines what type of property HHs can afford. Therefore, from a policy standpoint, reliable property price data are key to understanding the scale of affordability issues and identifying market failures leading to supply-demand mismatches in the housing sector.

House price data also bear key insights on countries' financial and macroeconomic stability. Disproportionally rising private property prices, or high price-to-income ratios, can be an early indication of imbalances and risks in the financial system (e.g., Anundsen et al. 2016; Drehmann & Juselius 2014). Decreasing house prices, in turn, may be associated with a decrease in HH wealth and a decline in consumption (e.g., Campbell & Coco 2007; Mian et al. 2017). If monitored over time, property prices can function as an early warning system of systemic banking stress or economic downturn. In addition, urban planners, private developers, economists, and policy makers depend on reliable property price data to update zoning regulations, formulate housing policies, and decide how to efficiently allocate scarce resources to support housing solutions where they are needed most.

Despite the importance of property price data for various stakeholders, comprehensive house price data are scarce in developed economies and virtually non-existent for emerging markets and developing economies (EMDEs). If available at all, residential property prices are mostly presented and published in indexed format, allowing to track changes over time; however, indices are not informative to understanding distributional aspects of prices, affordability issues, or the degree of price segments underserved in a market. When housing markets mature, properties' characteristics and attributes may be more stable over time. Hence, changes captured by price indices reflect price movements over a relatively constant set of properties in the market. However, in many EMDEs, there are dynamic changes to the type of properties being built by formal developers. Due to the nascency of markets, formal developers and builders in EMDEs are evolving to expand their portfolio to also cater to housing solutions to HHs with lower or even informal incomes. In such circumstances, aggregate price indices may not be fit for purpose and may suffer from biases due to rapidly changing underlying property types.

The wide gap of publicly available property price data in EMDEs is a key impediment to extending the understanding of housing markets widely available from developed economies to these markets. Insights on the determinants of property prices and their drivers over time are heavily researched subjects, but exclusively rely on studying this dynamic in developed economies. If comparable property price data were available in EMDEs, they could provide crucial insights on potential inefficiencies along the housing value chain. Analyzing property price data in emerging economies can also help, for instance, to point towards constraints in developers' ability to access finance, prohibitively high construction costs, or regulatory bottlenecks in land acquisition and titling. In addition, property price data in EMDEs can also help point towards affordability issues and potential ramifications on HH consumption and spending. The width of these challenges is often difficult to grasp as the housing value chain is very interconnected and complex, varying across contexts and countries. Despite the immense importance of understanding these challenges and identifying how to mitigate housing market failures, the empirical literature in this area is highly underdeveloped for EMDEs. This can be primarily attributed to a dearth of robust property price data and associated analysis in these markets.

This paper pursues a two-fold approach to address the gap: First, addressing the information vacuum in EMDEs regarding property price data, this paper presents a novel approach to collecting house price data in data-scarce environments common in EMDEs. Instead of relying on survey data or obtaining property price data from various official and unofficial sources, we scrape available price information from EMDEs' listing websites where private properties are listed for sale. In doing so, we largely constrain our analysis to the formal housing market and disregard informal housing not transacted in online markets, such as the incremental self-building of houses typical in EMDEs. Our approach offers a first foray in providing the distribution of property price data otherwise nonexistent in EMDEs.

In the second part of this paper, we demonstrate the useability of distributional private property price data and present an approach to estimate the price of a "standard house". We refer to this price as the Basic House Price (BHP). This BHP allows for comparability of prices across EMDEs, standardizing the significant variation in the type of property across markets. Further, the BHP presents a standard measure that can be used in future research to monitor aspects of inclusiveness. When linked to other data such as, for instance, household income or housing finance data, progress on the Sustainable Development Goals (SDGs) may be monitored. The BHP also offers a standard measure to assess and compare housing affordability across population segments if paired with income data.

We present the web scraping approach, descriptive statistics, and the methodology for estimating the BHP exemplarily for five countries: Albania, Costa Rica, Morocco, Pakistan, and South Africa.[2] Overall, we collect residential property price data from over 200,000 online listings and present an approach to clean, analyze, and visualize these data. In this paper, we only demonstrate the applicability of the collected house price data for the price estimation of a defined housing unit. We leave for future research on how a standard house (BHP) price may contribute to understanding affordability in EMDEs.

This paper proceeds as follows: After this introduction, Section 2 discusses previous efforts in collecting and estimating house prices, which are mainly focused on developed economies. This section outlines how – if collected at all – private property prices are mostly available in an indexed format which only allows for monitoring of changes over time. Section 3 describes the data collection process for the five countries and discusses the prospects and pitfalls of a web scraping approach to obtain data for EMEDs. Section 4 defines a standard housing unit, which is the underlying concept for estimating the BHP. This section also rolls out how we apply a Random Forest (RF) model to estimate the BHP. Section 5 presents BHP results for the five markets, including estimations for the largest cities in these countries. Section 6 concludes and outlines how the methodology we propose can be expanded to a more extensive set of EMDEs.

[2] Our ongoing research extends the collection and estimation of private property prices to over 60 EMDEs and will be presented in future papers.

2. Literature: Collecting and Estimating House Prices

Despite their importance, private property price data are not readily available for most countries and are particularly scarce for EMDEs. Further, the lack of standardization of property prices and the heavy reliance on indices to monitor changes over time make comparisons across countries cumbersome. Addressing this gap, the paper places itself at the crossing of three streams of literature: existing efforts to collect or collate data on residential property prices and to construct residential property price indices (Section 2.1); studies on determinants of house prices (Section 2.2); and a relatively new area of big data and machine learning approaches to estimate (determinants of) house prices (Section 2.3). Finally, we summarize the existing gap within this line of research and outline how we address it (Section 2.4).

2.1 Current Efforts to Collect House Price Data and Existing House Price Indices

Since housing plays a key role in the growth of many aspects of a country's economy including the development of the construction industry, job creation, and improving the living conditions of many HHs, governments have a great interest in understanding property prices and their developments over time. Most advanced economies' statistical offices or central banks, therefore, started collecting data on residential property prices in the 1970s (Knoll et al. 2017). Also, tax authorities, land registries, or real estate associations collect, hold, and sometimes even publish data on residential property prices in many advanced economies.[3] With the 2008-2009 global recession, which many scholars attributed to misalignments in housing and housing-related asset prices, the interest in dynamics in housing markets rose significantly (e.g., Goodhart & Hofmann 2008; Del Negro & Otrok 2007). Since then, several international organizations and central banks have increased efforts to develop global property price indices and collated real property price data for various predominantly developed economies to monitor macro-financial stability and price developments. One early methodology to track property prices is the Case-Shiller National Home Price Index which measures the value of residential real estate in major US metropolitan areas and serves as a blueprint for subsequent price indices in other developed countries.[4] Extending the collection of residential property prices to EMDEs has been slow, in either indexed or other forms. In the following section, we briefly discuss the most important sources.

First, a primary source for residential property prices, covering a relatively large number of countries, is provided by the Bank for International Settlements (BIS). BIS collects quarterly data on residential property prices for 60 countries, predominantly focusing on advanced economies. Property prices are harmonized as much as possible by BIS according to the recommendations outlined in the Handbook on Residential Property Price Indices (RPPIs), which summarizes best practices in how to calculate property price indices (European Union [EU] et al. 2013). As BIS compiles data from various sources, the data series differ from country to country, varying in frequency, type of property, covered area, priced unit, compilation method, or seasonal adjustment. In addition, while BIS publishes some actual prices, most data are only available in an indexed format, allowing for tracking aggregate price movements over time. BIS does not provide

[3] The UK Land Registry, for instance, makes house price data publicly available: https://www.gov.uk/guidance/about-the-price-paid-data.

[4] These price indices, including twenty cities, low-, medium-, and high-tier home price indices, condominium indices, and a U.S. national index, are now published as the S&P/CoreLogic/Case-Shiller Home Price Indices by Standard & Poor's.

insights on the distribution of property prices.[5] Despite these shortcomings, this database currently offers the most comprehensive data series on house prices.

Second, the International Comparison Program (ICP 2011), which collects prices for a range of goods and services that make up final consumption expenditure and gross capital formation, also captures housing expenditures. The ICP survey collects annual rental prices and dwelling stock data. Rents are either captured as actual or imputed rents (World Bank 2020). In the most recent ICP cycle (2017), participating economies collected rental data for 21 different dwelling types, ranging from one-bedroom apartments to single-family homes.

Third, a more regional-focused data source on property prices is provided by the Organization for Economic Co-operation and Development (OECD), which publishes nominal residential property price indices for OECD countries, as well as price-to-rent and price-to-income ratios.[6] The database particularly focuses on house price developments across regions and cities within countries to capture spatial price variation. For select countries, OECD also offers the number and value of housing transactions. While insightful for advanced economies, this database does not cover any emerging economies and mostly publishes indexed data to track price changes over time.

Fourth, institutions such as the International Monetary Fund (IMF) or the United States Federal Reserve Bank collate property price data from various national sources. IMF's Global Housing Watch platform, for instance, tracks developments in housing markets across the world on a quarterly basis.[7] The database collates property price data from different sources (e.g., BIS, European Central Bank, Federal Reserve, and national source) for 63 countries – mostly advanced economies – to assess valuation in housing markets. Further, it provides metrics such as price-to-rent and price-to-income ratios. Similarly, the Dallas Federal Reserve Bank's International House Price Database publishes quarterly house prices for 25 mostly developed economies by drawing on national public sources primarily from central banks, statistical offices, or other non-government organizations (Mack & Martínez-García 2011).[8] The data collected by these institutions are mostly for developed economies, and these institutions collate secondary data from a plethora of different sources, and primarily make data available in an indexed format.

Fifth, in collecting and analyzing actual house prices *across* emerging economies, the Center for Affordable Housing Finance in Africa (CAHF) is unique in its efforts. It systematically collects house prices for African countries by surveying local housing experts on the cost and size of the *cheapest* house built by a private developer. Figures are published in CAHF's annual housing finance yearbook, covering the last decade. CAHF's approach also turns the conversation on house prices away from mean or aggregate measures and to the lower tail of the formal market. From the perspective of policy dialog on affordable housing, this approach may be more appropriate. However, the usability of the data for policy purposes suffers partly because i) the price point provided represents only the extremely lower end of the formally developed new housing units and ii) is not paired with information regarding the quantity supplied at or near this price range.

[5] https://www.bis.org/statistics/pp_detailed.htm
[6] https://data.oecd.org/price/housing-prices.htm
[7] https://www.imf.org/external/research/housing/index.htm
[8] https://www.dallasfed.org/institute/houseprice

Finally, in recent years, crowd-sourced platforms such as Numbeo,[9] which rely on user inputs on property prices in various locations around the world, have added their own house price index along with publicly available per square foot price ranges for properties within the city center and outside the city center. These platforms add more distributional aspects to the average house prices and point predictions published by other indices, but suffer from the reliability of the self-reported data.

Despite the apparent issues in comparability across countries, the listed sources are the most comprehensive databases on property prices currently available for a larger set of countries. Therefore, many papers draw on these indices to conduct country or region-specific analyses on property price developments over time (e.g., Girouard et al. 2006; Igan & Loungani 2012; Yoshino & Helble 2016). One of the earliest systematic presentations of house prices is a historical time series data set of nominal residential property prices in 13 advanced economies by Borio et al. (1994). Some studies that provide comparative assessments combine the data sources outlined above or enhance them with some primary data collection on some additional countries that are not yet covered by the indices (e.g., Deghi et al. 2020).

2.2 Determinants of House Prices

The volume of research on the housing market, particularly estimating its impact on real economic activity, has experienced a steep influx after the global financial crisis in 2008–2009. Most studies in this realm investigate the various channels through which housing and house prices affect macroeconomic and financial outcomes, particularly as housing bubbles are associated with significant output losses (e.g., Catte et al. 2004; IMF 2008; Jordà et al. 2015). Single-country studies on house prices and house price developments mainly focus on developed economies, particularly OECD countries, EU countries, and the United States or Canada (e.g., Alter & Mahoney 2021; Davis & Heathcote 2005; Knoll et al. 2017; Philiponnet & Turrini 2017).

Most studies investigate determinants of house prices over time. Jordà et al. (2016), for instance, have gathered time series data on disaggregated bank credit for 17 advanced economies since 1870. With this historical data for the total value of the residential housing stock (structures and land), the authors relate household mortgage debt to asset values, showing that the rise in mortgage credit has financed a substantial expansion of home ownership from about 40 percent in 1950 to 60 percent in the 2000s. Similarly, Knoll et al. (2017) assess how house prices have evolved over time for 14 advanced economies, gathering historical house price data to estimate what drives changes in house prices. The authors show that changes in house prices are largely attributed to changes in land prices. This finding is corroborated by others who also attribute rising property prices to sharp increases in residential land prices, while construction costs have remained relatively stable over time (e.g., Glaeser & Ward 2009; Gyourko et al. 2013). In major metropolitan areas, it is not uncommon for the cost of land to exceed 40 percent of total property price; in extreme cases, like San Francisco, the share can stretch to as much as 80 percent (McKinsey Global Institute 2014).

Gao et al. (2019) dissect property features into two groups when predicting house prices: non-geographical features, such as the number of bedrooms and floor space area, and geographical features, such as the distance to the city center and the quality of nearby schools. This is also documented by Gröbel and

[9] https://www.numbeo.com/property-investment/

Thomschke (2018) who show that housing prices are largely determined by the physical location of the property. In addition, the number of bedrooms and the size of a private property are consistently found to be positively related to the property price (e.g., Fletcher et al. 2000; Garrod and Willis, 1992; Rodriguez and Sirmans, 1994). Other attributes studied include crime rates (e.g., Ceccato & Wilhelmsson 2020), or proximity to transportation (e.g., Zhang et al. 2021; Zong & Li 2016). Other authors estimate that house prices have particularly increased since the financial crisis in 2008 due to the rise of economic activity paired with unusually low mortgage interest rates in most advanced economies (Claessens & Schanz 2019). Also, price changes in major cities are hypothesized to be driven by institutional investors trying to find high yields or safe assets in a low-interest rate environment (IMF 2018; Gauder et al. 2014).

While there is a plethora of literature on property price determinants in developed economies, studies on EMDEs are scarce. Some single-country studies focusing on EMDEs are analyzing existing house price data that are published by commercial banks such as e.g., Absa, Standard Bank, and First National Bank for South Africa (e.g., Balcilar et al. 2011; Luüs 2005). Other authors collect their own data by either surveying real estate agencies to estimate the relative importance of housing attributes to house prices (Owusu-Manu et al. (2019) for Ghana), by surveying developers (Libertun de Duren (2018) for peri-urban areas in Brazil and Mexico), or by conducting a household survey to collect data on housing costs (Uwayezu & de Vries (2020) for Kigali city in Rwanda). High property prices in EMDEs are often attributed to prohibitively high building costs due to the need to import materials, the shortage of local skills, and the absence of financial mechanisms that allow for materials to be bought in bulk (e.g., Gardner & Pienaar 2019). While unique in their efforts to shed some light on the housing market in emerging economies, these studies provide only a snapshot of the housing market of one country (or a handful of countries) – often with a regional focus or a focus on the biggest cities.

2.3 Big Data Approaches and Machine Learning for Private Property Price Estimation

In addition to traditional approaches of data collection of private property prices discussed in the previous section, in recent years and with the gaining momentum of big data and machine learning in economics, more studies started to gather property price data from online listing websites. While less than a decade ago, most private properties were listed for sale in local newspapers or with private realtors, today, much of the listing activity has moved to websites concentrating on housing advertisements.

Analyses that draw on property price data collected from listing websites allow for fine-grained spatial and temporal assessments of the entire housing market. Further, big data approaches to private property prices will enable one to investigate a particular housing market in more detail or add distributional aspects to the mostly averaged house prices made available by indices discussed in Section 2.1.

A predecessor of web scraping approaches to collect property price data includes Kim's (2007) study on Vietnam. The author manually collated over 5,000 observations on property prices and property attributes drawing on classified advertisements in Vietnam's most prominent newspaper. Applying a hedonic price model, Kim assesses the price differences between Hanoi and Ho Chi Minh City to investigate the impact of social norms on property prices. Over time and with the increased penetration of property listing websites, private property price collection efforts have transitioned to online listings where data collection can be automated. Anenberg & Laufer (2017), for instance, use listing information to construct a new house price index to monitor house price developments in the US. Using property listings, the authors construct a new repeat-sales house price index that describes house values at the contract date when the price is determined rather than the closing date when the property is transferred. Other big data price collection efforts include, for instance, scraping of online listings in Great Britain (e.g., Rae 2015), the US (e.g., Boeing et al. 2021), the Netherlands (ten Bosch & Windmeijer 2014), Türkiye (Keskin & Watkins 2017), Japan (Sadayuki 2018), or China (Hu et al. 2019; He et al. 2019; Wang et al. 2020).

In their data collection efforts, most authors focus exclusively on a localized housing market (i.e., a particular region, city, or neighborhood) in developed countries for which well-structured property listing websites with a plethora of private properties listed for sale are available. Additionally, while very comprehensive in scope, most efforts of web scraping of private property prices are centered on developed markets. Similar approaches in EMDEs, particularly in low-income economies, are scarce. Notable exceptions are, for instance, Gnagey and Tans (2018), who collate a data set of over 64,000 properties in 2016 from listing websites to estimate house prices in Indonesia. The authors find that desirable housing attributes, structural quality, advantageous location on major thoroughfares, and secure land tenure increase property asking prices.

Almost all studies that collect price data from online listing websites focus on only one or few markets within a particular region. One notable exception is a recent *HouseLev* database project that assembled house prices for 40 countries, mainly European and advanced economies, including some emerging economies such as Türkiye or the Russian Federation (Bricongne et al. 2019). The authors do not solely rely on web scraping for all 40 countries. They instead relate to national accounts data and implement web scraping as a "fallback methodology" in case of missing data. As the authors use both methods, national accounts as well as web scraping, for a sub-sample of European countries, they can compute the median level of estimated upward bias arising from the use of listed rather than transaction prices, which is then applied as a correction factor to improve comparability of price level data obtained with the two methods (Bricongne et al. 2019: 6). *HouseLev,* to the best of our knowledge, is the most comprehensive web scraping project of private property prices, primarily focused on developed economies.

Advanced price estimation techniques have also evolved with the increased usage of big data approaches to collecting property price data from listing websites. Traditionally, the hedonic price model, which draws on Lancaster's consumer theory, has long been the predominant model to estimate property prices (Lancaster 1966; Rosen 1974). Property prices are modeled in multiple regression analysis, assessing the association between property price and several hedonic attributes through parametric estimation (Oladunni & Sharma 2016). Attributes frequently applied in hedonic price models include, for instance, number of bathrooms, number of bedrooms, area size, neighborhood, or accessibility of the property (e.g., Borba & Dentinho 2016; Can 1992; Krol 2013). While very simple in their interpretation, hedonic price models require the fulfillment of strong model assumptions, including functional form of the conventional hedonic pricing model, homoscedasticity, independence, and the absence of multicollinearity (e.g., Anderson 2000; Pérez-Rave et al. 2019). [10]

In recent years, the applicability of alternate methods to the hedonic price estimation has expanded and machine learning (ML) has emerged as an alternative to predicting house prices (Borde et al. 2017; Čeh et

[10] For a critical overview of the different prediction algorithms commonly used for house price predictions, see Montero & Fernández-Aviles (2018).

al. 2018; Fan et al. 2006; Mullainathan & Spiess, 2017; Pérez-Rave et al. 2019; Truong et al. 2020; Yan & Zong 2020). Within that realm, Fan et al. (2006) constitute one of the earliest contributions that move beyond hedonic price models to predict property prices. Applying a decision tree technique, the authors explore the relationship between house prices and housing characteristics, which aided the determination of the most important variables for price predictions.

While ML techniques are comparatively weak in inference, they have strong predictive power, manage to fit complex data, are very flexible in assumptions on functional form without overfitting, and work well in out-of-sample estimations (e.g., Athey 2018; Mullainathan & Spiess 2017). ML estimations such as random forest have become a suitable, and frequently applied alternative to hedonic price estimates, particularly for property price estimation. While RF and other decision tree-based models also rely on model assumptions, they are better at modeling non-linear relationships compared to simple, multi-linear regression.

Authors applying ML to price estimations mostly focus on narrowly defined housing markets in developed economies such as Ljubljana, Slovenia (e.g., Čeh et al. 2018), Gangnam, Republic of Korea (e.g., Hong et al. 2020), London, Great Britain (e.g., Levantesi & Piscopo 2020), Arlington County, USA (e.g., Wang & Wu 2018) or housing markets in upper-middle income economies such as Mamak District, Ankara, Türkiye (Yilmazer & Kocaman 2020), Petaling, Jaya, Selangor, Malaysia (Mohd et al. 2019), or St. Petersburg, Russian Federation (Antipov & Pokryshevskaya 2012). In assessing the housing sector, many of these authors contrast the predictive performance of ML algorithms with standard regression techniques. Across the board, the authors find that RF (significantly) outperforms parametric estimation techniques in terms of accuracy and predictive power.

2.4 Data Gap: EMDEs Are Largely Absent in Property Price Analyses

This overview on existing studies and data sources within the realm of house prices points to five major gaps that we try to address with this paper:

First, there is a striking data gap in the availability of house prices, particularly for EMDEs. Most existing property price compilation efforts concentrate on developed economies, publish data only in indexed format, and do not report underlying actual house prices. This may be attributed to the fact that national sources, such as central banks or statistical offices on which these indices base their data, do not collect, report, or publish property price data. Further, as underlying data to these indices are very country- and context-specific, they fit the purpose of monitoring price changes over time within a specific country but do not facilitate cross-country assessments. Some countries, for instance, only consider prices for family homes in the capital while other countries use flats in urban areas for the index. The same applies for prices: some report the transaction prices while others draw on listed prices, while yet others average prices (cf. BIS database; Mack & Martínez-García 2011). Mack & Martínez-García (2011), who collate publicly available national sources to build a database of (nominal and real) house prices for developed economies, acknowledge this flaw outlining that the main contribution of their database is "sorting out the existing data by country, selecting the most *similar* series and documenting the differences across countries to clarify the extent to which international sources can be made comparable for empirical analysis purposes" (Mack & Martínez-García 2011: 3). Achieving comparability across countries with the existing data sources is almost impossible.

Second, while price indices present equilibrium outcomes of housing markets, they do not cover details about, broadly, the quantity of housing. They often only include a particular type of housing for which prices are tracked. Whereas in high-income economies, the latter may remain relatively stable in the short term, in EMDEs, with rapidly expanding formal housing markets, quantity and type of housing are important elements to capture. They provide context to changes in prices as the sample over which prices are indexed changes, and as price and quantity and type of housing supplied are highly interrelated. Also, they have important policy implications in the context of markets' ability to supply homes for different market segments, and formal developers' ability to go reach lower income groups. The measurement of these dimensions of the housing markets is absent across EMDEs.

Third, primary data collection for actual, non-indexed house prices is still somewhat limited and, if available, almost exclusively covers advanced economies. With a few notable exceptions, there is a severe lack of contributions in the literature on property price estimations in EMDEs. Property price data for EMDEs are virtually non-existent – both for within-country assessments, and even more so for cross-country comparison. Most studies on house prices obtain data from readily available sources such as land registries, real estate agencies, or commercial banks, or tap into established indices. Given the significant effort to collect original data on house prices, there are very limited efforts. Since the scope and focus of these studies differ or as they purely rely on price indices, comparing property prices across studies is not feasible.

Fourth, efforts to investigate property prices in EMDEs mainly converge to analyze the determinants of *mean* house prices. Distributional efforts in property price collection for different income segments within emerging economies are largely absent. CAHF is unique in its effort to approach the house price estimation from the perspective of low-cost developers. Yet, CAHF takes it to the other extreme. It only collects the *cheapest* price of a house built by a formal developer in African countries and does not factor in otherwise transacted housing units in the formal housing market. While insightful, this approach does not allow pricing the entire housing market, offering an understanding of the quantity of "affordable" houses available to different income segments.

Lastly, studies assessing the historical developments of private property prices are concerned with measuring financial (in)stability, which they attribute to distorted household mortgage debt to asset values ratios. A myriad of studies estimating house prices in developed economies were published after the collapse of the housing bubble and the resulting financial crisis in 2008/2009 mainly concerned with estimating how to identify housing bubbles in the first place. An examination of property prices from the perspective of affordability and demand-supply mismatches for different income segments is absent. In addition, studies assessing house price data usually draw on different methodologies to estimate property prices and rely on varying data sources. Hence property price estimations are not comparable across studies and scholars have only recently started to use big data approaches to collect actual property price data for a larger number of economies. Yet, their focus mostly remains on developed countries.

To fill these gaps, our paper extends the novel approaches in collecting house prices through a web scraping approach to emerging economies, thus addressing the substantial data gap in EMDEs. Further, this paper offers a methodology contributing to a more distributional understanding of private property prices in emerging economies. It also provides a comprehensive methodology to estimate a standard house price that allows for consistent price comparison across countries. These data can then facilitate the extension of the scope of the analysis to affordability assessments of property price data and the segmentation of the housing market – particularly focusing on EMDEs.

3. Data Collection and Processing: House Price Data in Emerging Economies

In this section, we outline how we collect house price data through a web scraping approach for five markets: Albania, Costa Rica, Morocco, Pakistan, and South Africa. We demonstrate how a big data approach, hitherto employed mainly in developed economies with good data quality, can also be extended to EMDEs to collect price data efficiently. We collected 200,000 unique property transactions for these five countries in an otherwise data-scare environment. The web scraped data reflect the entire housing market and complement the available indexed data that (mostly) report average property prices only.

We selected these five economies to cover different regions and factors in varying country contexts to highlight specificities of web scraping and data processing in EMDEs. These include, for instance, types of properties listed, unique forms of data entry specific to EMDEs, or cultural aspects. We do not strive for the representativeness of these five countries for all EMDEs but seek to exemplify the unique challenges of applying a web scraping approach in EMDEs. Nevertheless, transferring this approach to other EMDEs, especially those with lower data quality, will come with additional unique challenges (as discussed in more detail in Section 3.3).

3.1 Web Scraping House Prices in EMDEs

The transaction price would be the ideal source to obtain comparable property price data. Typically, these data can be found in land registries or tax authorities, collated from real estate agencies, collected through online surveys, or obtained through appraisals or valuations as part of the mortgage process. However, none of these sources are feasible for automated data collection in EMDEs as the various institutions holding these price data do not yet have a standardized way of collecting, publishing, or even digitizing them. In markets characterized by lax regulation or enforcement, transacting parties may under-declare property prices to avoid negative ramifications with respect to paying additional registration costs or taxes.

We opted to collect property price data for EMDEs through a web scraping approach of real estate websites. While not yielding transaction prices, obtaining listing prices of formal properties is a viable alternative to gathering price data in EMDEs, where data are otherwise non-existent. At least in the context of developed markets, strong evidence exists that listing prices are correlated and a good leading indicator for transaction prices (e.g., Ardila et al. 2021; Anenberg & Laufer 2017; Lyons 2019). For each economy, we scrape property prices and additional data points for all available listings, capturing, to the extent possible, location aspects. Scraping unique property transactions has several advantages: first, they allow us to provide actual property price data of an entire housing market in EMDEs, facilitating analysis beyond aggregated or indexed data. Particularly in EMDEs, we expect significant differences in prices between the biggest business city and rural areas, and more considerable skewness in data even within cities. Second, by web scraping property prices of the entire formal housing market, we can analyze sub-markets in greater detail. Third, collecting house price data from the entire formal, online housing market allows us to also capture the quantity of housing available at different price segments and can, therefore, provide an overview on the distributional aspects of the housing market within a given economy. These distributional price data are very relevant for additional analysis as they can, for instance, be paired with household-level income data for affordability assessments. Finally, an overview of the entire housing market can reveal supply-demand mismatches particularly regarding in which price segment formal housing market activity is generally low or absent altogether.

We start the web scraping process by identifying the most up-to-date and complete websites that list private property prices for sale in the five EMDEs. We identify up to three relevant listing websites per economy. Websites were selected based on the following aspects: (i) websites with the most comprehensive number of up-to-date listings, (ii) websites that offer broad ranges of properties and do not only cater to the luxurious segment (i.e., avoiding websites exclusively targeting expats etc.); (iii) websites that offer structured data entries on housing attributes including price and size. We limit ourselves to up to three websites since we notice considerable cross-postings in additional, usually less comprehensive, websites. We then scrape all residential properties that are listed at one point in time for sale on these websites along with all available housing features, including price, size, type (i.e., whether the property is an apartment or a house), location, number of bedrooms, number of bathrooms, and sometimes amenities such as garage, time of construction, number of floors, etc.

We extracted online listing data for the entire formal housing market of five EMDEs at one point in time, between April 2020 and August 2020. While this falls within the onset of the COVID-19 pandemic, insights on how house prices were impacted in EMDEs are qualitative and largely anecdotal. Commentary on the matter focuses on the affordability challenges for HHs rather than specifically on changes in property prices.[11] Beyond qualitative insights, comprehensive analyses on price changes due to the COVID-19 crisis are preliminary and focused mostly on developed economies (e.g., Pfeifer & Steurer 2020 for Vienna and London; or Bricongne et al. 2021 for the UK). While the results are not transferrable to EMDEs, they still offer some context into a largely under-researched area. On the impact of COVID-19 on housing markets in the UK, Bricongne et al. (2021) show that while the number of offers per week dropped during the first lockdown period, house prices did not change significantly (maximum of 2.6 percent increase) (Bricongne et al. 2021). Pfeifer & Steurer (2020) make a similar observation for the housing market in London, while showing that the housing market in Vienna follows an upward trend following the COVID-19 crisis. Despite the timing overlap, we cannot draw on existing literature to determine if bias may exist in our data, or, more importantly, the direction of the bias.

As a first effort to scrape private property prices in several EMDEs, we faced unique challenges compared to similar efforts in developed markets, such as issues pertaining to the number of observations available per website, the organization and reliability of data, measurement units provided, and the overall reliability of the websites. In many EMDEs, property listing websites are often not the primary source for transactions. Often, buyers and sellers revert to real estate agents and personal interactions. Yet, online platforms are becoming increasingly more popular for transacting goods and services, including properties. In Africa, for

[11] Some regional analyses in Latin America qualitatively point to the fact that while there was economic slowdown and increased investor uncertainty dampening growth in the short term, but also that the COVID-19 pandemic has shifted consumer preferences to larger properties with more outdoor space. Another analysis for India, for instance, reported that house prices have stagnated in 2020 / 2021, a trend attributed, among others, to the receding demand due to the COVID-19 pandemic (Reuters 2021).

instance, Jumia.com, which is an online platform combining an e-commerce marketplace, classified websites, and applications, is widely used across the continent. In Nigeria, Property Pro (formerly Jumia) is the number one property transaction website, covering about 65 percent of the Nigerian online real estate market (Nairametrics 2018).

Additionally, EMDEs' websites, particularly those that do not have a regional spread like Jumia.com in Africa, have limited formal standards regarding data entry. Many websites in emerging economies do not have consistent data on an array of property characteristics and provide somewhat limited information on the listed properties. Often, they only include some pictures of the property, the listed price, and a phone number through which the seller can be reached. Formal developers, who have started to also list newly built properties on online platforms (in addition to their own online or offline platforms), provide slightly more structured information on the transacted property. Yet, they are bound by the format of the online platform, which often only requires submission of property price and size. Few developers provide exhaustive property descriptions in free text format, which, if extracted, needs to be processed for data analysis through text mining. Hence, property data obtained through web scraping in emerging economies, in our experience, will not be as exhaustive in terms of obtaining different property features as found in publications on developed markets (cf. Section 2.2).

Furthermore, across websites in EMDEs, there is no standardized way to record the size of the property. Sometimes, the website does not provide the option to insert size information at all. In addition, user-provided information on size might not necessarily comply with the unit required by the platform (e.g., users insert square feet even though the platform requires square meters). The matter is even more complicated in economies where local measurement units for properties are used alongside more "standardized" measurements. South Asian websites (Nepal; Pakistan) allow for the insertion of different size units including Biga, Kattha, Dhur, Ropani, Aana, Paise, and Daam, alongside square feet and square meters. However, not all users consistently specify the measurement unit making data cleaning cumbersome. In addition, particularly for houses, size data can be somewhat muddled as it is unclear whether the plot size or the usable property size is indicated. To account for this difference, we distinguish apartments and houses (cf. Section 4.1) and, where available, use plot size for houses to also account for the value of the land, which – in some EMDEs – can be a significant portion of the property price (cf. Section 2.2).

Finally, in some EMDEs, listing websites do not specify whether a property is for rent or for sale. Usually, rental properties can easily be distinguished by relatively low prices. However, in some EMDEs, it is common to pay one year's rent upfront. In these instances, it is challenging to discern low sales prices from annual rents in cases where listings do not distinctly indicate sale versus rent. In addition, price data do not always include a currency marker, which is mainly problematic in countries where both euros and US dollars are used to transact properties in addition to local currencies.

When the aim is to estimate representative property prices of the housing stock in a country or region, scraping at least 0.5-1 percent of the number of HHs in that area is considered to be a large enough sample (cf. Bricongne et al. 2019). The same principle applies when the statistical population being analyzed is the universe of transacted properties in the market over a given period: in developed economies where residential property markets are formalized, it is possible to obtain the total number of transactions (the statistical “universe”) and thus sample appropriately to achieve representativeness. In EMDEs, by contrast, we expect most transactions to occur informally and outside of what is observable publicly. Through web scraping, we constrain our analysis to the formal housing market and to what is transacted online. With this approach, we are able to obtain house prices for the formal housing market but are unable to infer the degree the estimations apply to – in some EMDEs admittedly large – informal housing markets. We aim for representativeness of the formal housing market only, and to achieve this, we scrape entire websites to cover all available listings. Despite these efforts, we encounter smaller sample sizes in some countries, which are likely to stretch margins of error (cf. Section 3.3). Annex 1 provides an overview of the sample coverage and percentage of formal households scraped.

3.2 Data Processing: Data Cleaning and Outlier Removal

As with any data set that is obtained from user-inserted data, the scraped data is prone to incorrect, inconsistent, or missing information. Most online listing platforms do not run quality checks on the listed properties or require fully populated identification of property features. Preparing the data for analysis, we diligently cleaned the web scraped data removing data entry errors, duplicates, and outliers. Given the issues outlined in the previous section, which are inherent to EMDEs, data cleaning is more tedious and time-consuming than for more structured data likely to be obtained in developed economies. We describe the data processing steps in detail below, illustrating descriptive statistics of the various stages of the process (Table 1).

Duplicates

The first step in processing the data is identifying and flagging repeat data and duplicates. These mostly arise for two reasons: first, many EMDEs’ property websites allow for the re-submission of the same property within some days’ interval. Realtors mainly use this option to restore the property at the top of the search results list to improve visibility on the website. Second, duplicates may also arise because of cross-listing of properties across different platforms. Ideally, we want to create a data set that removes both occurrences. Hence, we deduplicate the data set to obtain what we call the *original data set*, dropping all exact duplicates that either have the same listing ID or that include the same title and description. Typically, properties with the same title and the same description are a clear indication for a repeat entry of a property on the same website. The title and description of properties, however, might bear similarities in instances where multiple, newly built apartments are advertised within the same complex. Also, in these cases, property features such as price, size, address, or number of bedrooms might be identical while referring to unique listings. Retaining these observations in the data set, we only remove *exact* duplicates with the exact values on price, size, bedroom, title, and description. A downside of this approach is that we run into the risk of keeping observations in the data set where the title or description has been slightly altered during the re-submission of the same property listing to the website.

As we assume that we will retain some duplicates in the data set, we also perform a more rigorous de-duplication where we remove all data that could potentially constitute a duplicate to understand how this alters our estimations. In this stringent outlier removal process, we remove all observations that have the same value on available property features only (price, type, bedroom, bathroom, and city) and disregard the title and the description of the property. While we note that this procedure is highly likely to also remove observations that are in fact unique but share the same property features, we perform this robustness test to ensure that repeat data do not drive estimated property prices. More sophisticated duplicate removal would include the use of text analysis techniques to understand the extent of similarity of the title or description of the property to remove those observations that have only been slightly changed during re-submission. Given the significant time effort of this technique, we opted for the more stringent data removal as a robustness test for price estimations (Annex 2).

Data Filtering

Next, we filter the data by excluding scraped data that are clearly not residential properties. These include storing units, garages, parking lots, undeveloped or agricultural land, or commercial properties. In addition, we truncate the data on price and size to exclude data entry errors and rental data. These include, for instance, rental data likely erroneously captured as sales price, particularly for those properties that include yearly rentals, spam or negotiable listings often detectable by “1” entered as the sales prices, and random data entries on square meter data. We apply a direct data filter to remove obvious errors and undesired data to obtain the *truncated data set*. We assume that all observations of below 9 square meters (sqm) and above 3,000 sqm are either data entry errors or properties that cannot be considered residential properties (e.g., storing units; large farmland). Also, we assume that properties of less than 9 sqm are not habitable for one person, aligning with the definition of the UN (UN-Habitat 2007). Regarding the truncation on price data, we remove any properties below 5,000 US dollars and above 50 million US dollars to account for data entry errors, rental prices that are accidentally listed as sales, as well as – a very typical feature in EMDEs – entire apartment complexes that are sold in bulk as an investment project. The major issue with apartment complexes or several apartment units being sold in bulk is the mismatch between the size and price data. Often, the price reflects the price of the overall apartment complex while the size reflects that of a single unit. Since it is often impossible to infer the actual per unit price and size, we exclude these properties to avoid distortion. While we apply a context-driven data filter to maintain as many observations as feasible in the *truncated data set*, we also apply a more rigorous winsorization to the data, common in large data sets such as ours (e.g., Bricongne et al. 2021 *for HouseLev Data*). We remove the first and 99th percentile of price and size and outline how this winsorization alters summary statistics (cf. Annex 3).

Outlier Removal

Having obtained the *truncated data set*, we perform additional outlier removal to ensure that skewed data do not drive estimations. Heavily, positively skewed property price data seem to be particularly acute in EMDEs where very luxurious properties catering to expatriates or foreign investors are transacted. To avoid analyzing severely skewed data, we employed two different approaches:

First, we right censored the data to remove luxurious residential properties that are not targeted at the local housing market. In doing so, we use Numbeo, a crowd-sourced global data platform that reports consumer prices, including private property prices in most countries’ largest cities. As Numbeo data is likely to be dominated by a bias towards data entry from higher-income individuals (with internet access), we consider Numbeo’s data maximum as the “true” maximum. Hence, we consider properties within the *truncated data set* that exceed the *maximum* per square meter price reported in Numbeo as an outlier. Hence, we obtain the *right censored data set*.

Second, to avoid that outliers at both tails of the distribution are distorting our estimations, we perform multivariate outlier removal on the *truncated data set* based on the robust Mahalanobis distance of each observation in the sample.[12] With this outlier detection technique, we remove outliers throughout the entire distribution, but mostly concentrated on the left- and right-hand tail.

Given the numerous data quality issues outlined above, we consider the second avenue of outlier removal, the more restrictive technique, most appropriate for our purposes and hence use the multivariate outlier removal technique to obtain the *final data set*. All other data sets are contained for robustness check purposes and to illustrate the data processing only (cf. Table 1). Price estimations and predictions are only performed on the *final data set*.

Data Sets Illustrating Data Processing

Table 1 summarizes the property price data of the five economies for the different stages of the data processing. In the *original* and *truncated* data set, the means of price, size, and per square meter price are (much) greater than their medians as the distribution is positively (and in some cases strongly) skewed by outliers.

While more robust statistics such as median and the interquartile range (IQR) stay relatively consistent across data sets, the standard deviation and mean drop significantly from the *original* data set to the *final* data set. This pattern remains constant throughout the five countries and provides some suggestion that the outlier removal process, while comprehensive on distance metrics, does not significantly alter the balance of the right and left tail and the order of the distribution.

In Albania, the difference in standard deviation between the *original data set* and the *final data set* is stark despite the relatively low number of outliers being removed. In Morocco, the *right censored data set* is the same as *the truncated data set* as the maximum price listed in Numbeo is smaller than the maximum price in the truncated data set, hence, no right censoring is applied here. In South Africa, the *truncated data set* – particularly if compared to the other four countries – excludes a relatively larger share of data entry errors and potential rental data. This might be attributed to large farms being included for sale on the website we used for South Africa (property24.com). In Costa Rica, the multivariate outlier removal technique detected particularly high-end, luxurious properties.

[12] We are applying the *smultiv* command in Stata, which has consistently proved to outperform *mcd*, an alternative robust estimator for outlier detection (e.g., Verardi & McCathie 2012.).

Table 1. Illustration of Successive Stages of Data Processing

Albania

Descriptive Statistic		*Original* Data Set	*Truncated* Data Set	*Right Censored* Data Set	*Final* Data Set
Number of observations		3,389	3,376	3,303	2,709
Price (USD)	Median	97,619.05	97,619.05	97,619.05	89,285.71
Price (USD)	Mean	137,498.50	137,483.8	137,155.6	99,599.84
Square meter	Median	97,00	97,00	98,00	92.00
Square meter	Mean	116.73	115.84	116.98	92.40
Price per square meter	Median	1,046.57	1,047.62	1,047.62	1,035.87
Price per square meter	Mean	1,227.50	1,219.54	1,219.54	1,075.40
Price per square meter	IQR	508.32	505.95	505.95	423.66
Price per square meter	SD	2,886.30	2,804.12	2,804.12	326.74
Number of bedrooms	Mean	1.98	1.98	2.00	1.78

Costa Rica

Descriptive Statistic		*Original* Data Set	*Truncated* Data Set	*Right Censored* Data Set	*Final* Data Set
Number of observations		10,376	10,202	10,159	7,551
Price (USD)	Median	200,000.00	200,000.00	200,000.00	175,000
Price (USD)	Mean	293,751.60	295,211.70	295,126.70	185,570.10
Square meter	Median	181,00	181,00	181,00	150,00
Square meter	Mean	347.11	234.17	234.17	160.71
Price per square meter	Median	1,156.72	1,162.28	1,162.28	1,273.90
Price per square meter	Mean	2,744.14	1393,17	1393,17	1,138
Price per square meter	IQR	673.14	665.61	665.61	569.76
Price per square meter	SD	80,189.03	3,869.17	3,869.17	543.29
Number of bedrooms	Mean[13]	.	.	.	.

Morocco

Descriptive Statistic		*Original* Data Set	*Truncated* Data Set	*Right Censored* Data Set	*Final* Data Set
Number of observations		10,734	10,479	10,479	8,673
Price (USD)	Median	104,395.60	105,494.50	105,494.50	93,406.59
Price (USD)	Mean	271,361.30	234,233.80	234,233.80	117,836.70
Square meter	Median	95,00	95,00	95,00	88,00
Square meter	Mean	188,59	133.68	133.68	95.47

Morocco (cont'd)

Descriptive Statistic		*Original* Data Set	*Truncated* Data Set	*Right Censored* Data Set	*Final* Data Set
Price per square meter	Median	1,119.25	1,117.21	1,117.21	1,039.27
Price per square meter	Mean	3,071.55	1,677.39	1,677.39	1,178.07
Price per square meter	IQR	932.10	903.46	903.46	776.44
Price per square meter	SD	33,458.45	7,570.71	7,570.71	557.07
Number of bedrooms	Mean	2.65	2.64	2.64	2.44

[13] The number of bedrooms in Costa Rica is missing because it was not available consistently from scraped websites.

Pakistan

Descriptive Statistic		*Original*	*Truncated*	*Right*	*Final*
		Data Set	Data Set	*Censored*	Data Set
				Data Set	
Number of observations		107,652	107,524	107,446	75,233
Price (USD)	Median	83,934.34	83,934.34	83,934.34	60,930.12
Price (USD)	Mean	156,756.10	154,420.40	154,487.40	68,011.17
Square meter	Median	151.76	151.76	151.76	126.47
Square meter	Mean	955.86	220.93	220.93	134.53
Price per square meter	Median	535.38	535.38	535.38	491.63
Price per square meter	Mean	634.36	633.16	633.16	507.60
Price per square meter	IQR	344.14	344.14	344.14	245.81
Price per square meter	SD	652.22	480.86	480.86	205.29
Number of bedrooms	Mean	3.86	3.86	3.86	3.33

South Africa

Descriptive Statistic		*Original*	*Truncated*	*Right*	*Final*
		Data Set	Data Set	*Censored*	Data Set
				Data Set	
Number of observations		91,904	88,374	70,155	42,369
Price (USD)	Median	95,588.23	93,286.45	102,301.8	68,734.02
Price (USD)	Mean	155,699.4	147,055	157,654.9	87,676.75
Square meter	Median	350	313	312	110
Square meter	Mean	2,081.1	555.01	554.09	193.98
Price per square meter	Median	360.38	387.17	387.17	580.06
Price per square meter	Mean	887.42	629.44	629.44	734.70
Price per square meter	IQR	614.62	624.15	624.15	620.48
Price per square meter	SD	9,302.60	853.02	853.02	654.53
Number of bedrooms	Mean	3.04	2.99	3.0	2.4

Note: The *original* data set contains the original set of all listings. The *truncated* data set retains listings that contain sale prices, size data, and whether the property is an apartment or house and truncates the data based on sqm<9 or sqm >3,000 and Price < $US 5,000 or Price > $US 5,000,000. The *right censored* data set retains unique listings that contain sale prices, size data, and whether the property is an apartment or house, truncates the data based on sqm<9 or sqm >3,000 and Price < $US 5,000 or Price > $US 5,000,000, and additionally removes the most extreme price data points on the upper end of the spectrum. The *final* data set retains thorough listings with reasonable values for price, size, and per square meter price, cleaned through multivariate outlier removal. SD= Standard Deviation; IQR=Interquartile range

3.3 Limitations

While applying a web scraping approach to obtain price data in EMDEs is a very cost-effective and efficient way to collect data, particularly compared to conducting expensive surveys, there are several limitations that pertain particularly in the context of EMDEs.

To start with, the web scraping approach does not necessarily yield data that are representative of all properties in the market as we are only able to capture properties of sellers with access to internet and who are able and willing to post their property online. By the same token, accessing house prices on real estate listing websites in emerging economies requires buyers to have access to these online listings. This might not always be the case, especially in lower-income segments of a given market. Particularly in developing areas, information density is low and might lead to data blind zones (Li et al. 2019). Recent research shows that online platforms used for home sales, even in developed markets, may reproduce and even intensify existing forms of inequality within cities (Boeing et al. 2021; Angelo & Vormann 2018).

While internet access is less of a concern in the five countries we outline here (cf. Annex 1), expanding the methodology to other countries might become problematic. In Burundi, for instance, only 5 percent of the overall population use the internet either via computer, mobile phone or other digital devices (World Bank 2022). In comparison, in Brazil, close to 74 percent of the population use the internet (World Bank 2022). In countries with relatively low internet penetration rates, HHs might adhere to alternative pathways to buy properties: personal interactions with real estate agents, classified ads in newspapers or through informal, personal interaction. Hence, the web scraping approach might not be suitable to capture local property markets where online advertisements are not frequently used and might, therefore, weaken the generalization of the results to localized markets.

Second, the price data collected are concentrated in countries' biggest business cities and urban centers. This is not surprising, since urban centers are the place where most new housing units are being built, responding to the accelerated urbanization rates currently observed in EMDEs. Further, urban dwellers are more likely to formally transact their property and to use online sources to sell or buy properties. Given the diversity of urbanization across EMDEs, the level of geographical disaggregation differs significantly. Disaggregated data for geographical areas beyond the major business city might not be sufficiently large to provide price estimations beyond the largest urban area. Due to data limitations beyond the biggest business cities and the absence of location markers on many housing listing websites in EMDEs, highly complex and spatially heterogenous housing markets cannot fully be delineated.

Third, the data listed on real estate websites include newly developed properties and the secondary housing market, which might bias house price estimations. In addition, some new housing developments are sold as investment projects, often tailored towards foreign investors. These properties are usually sold in bulk, i.e., entire apartment complexes containing several apartment units. Accounting for this potential bias, we conduct careful, multivariate outlier removal. In addition, we differentiate between property types (apartments and houses) and provide distinct price estimates for both property types.

Fourth, the final transaction price is likely to be different from the listed price, which often appears to be the price ceiling that precludes the possibility of sales at higher prices (Horowitz 1992). Furthermore, the listed prices advertised online represent the user-inserted price, which could include either the appraised values from some third party such as a tax assessor, or the self-appraised property values of homeowners. Regarding the latter, several studies have pointed towards a large variance of self-appraised values which in large enough samples like ours, positive and negative errors tend to cancel each other out (e.g., Follain & Malpezzi 1981; Goodman & Itter 1992). The difference between the transaction and the listed property price is dependent on multiple factors including the overall state of the housing market, the demand for housing, the time the property remains on the market before willing and able buyers come forward, as well as cultural aspects pertaining to e.g., negotiation. Despite these issues, in the absence of transaction price data sets, listing prices offer a good proxy to estimate the state of the housing market as researchers have consistently found rather low deviations between listed price and transaction price (Arnott 2009; McGreal & Taltavull de La Paz 2013; Haurin et al. 2010).

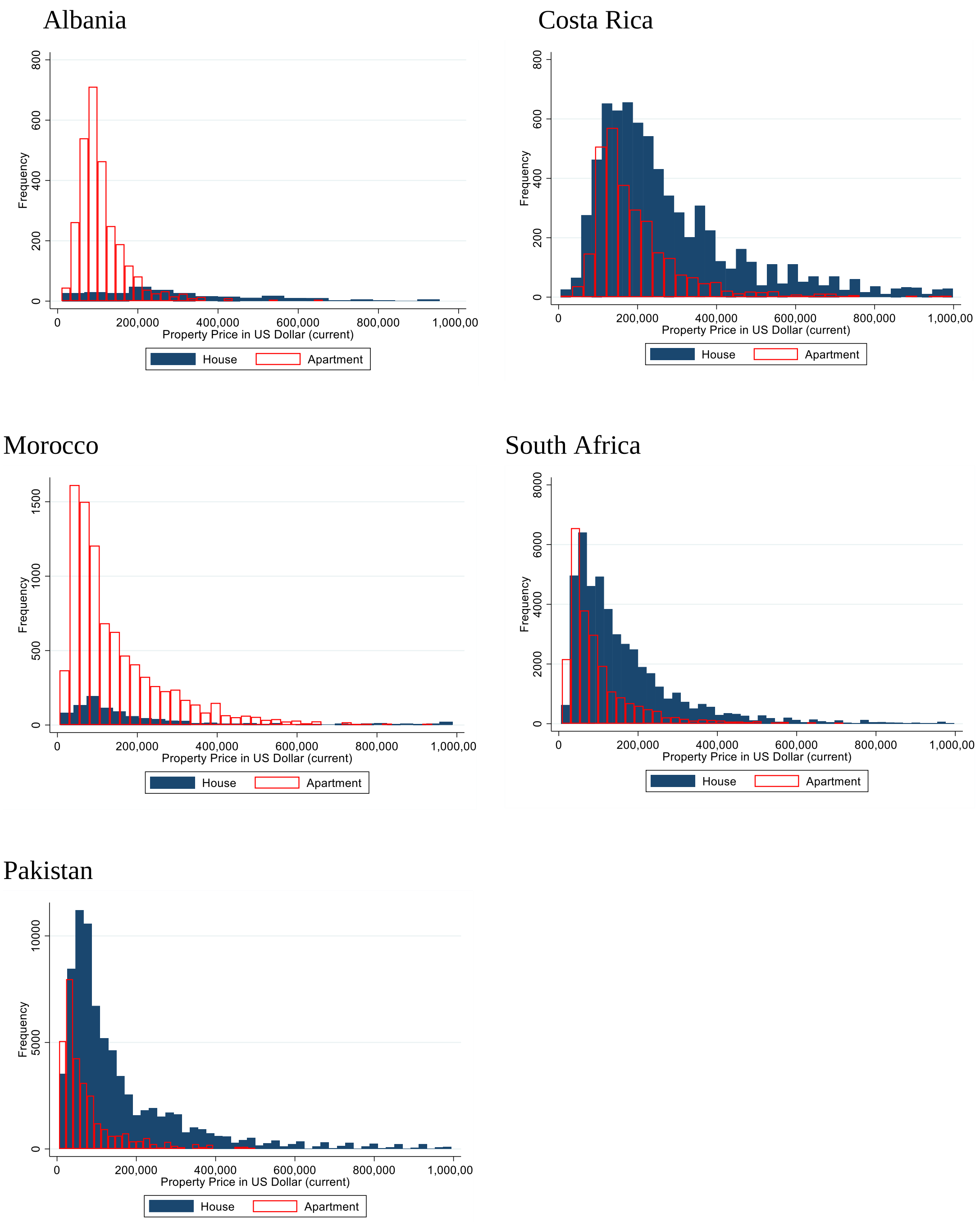
3.4 Descriptive Statistics

Across the *final data set* of the five EMDEs in our sample, we observe different patterns in terms of availability of apartments versus single family houses (Figure 1). While in Costa Rica, Pakistan, and South Africa, the number of apartments and houses are well distributed, Albania and Morocco have many more apartments than houses available within the data. In South Africa and Costa Rica, the right-hand side of the distribution is dominated by rather expensive single-family houses and only few, expensive apartments. Similar patterns are observable in the price-size relationship (Figure 2). While in Pakistan the price and size differences between houses and apartments are marginal, South Africa – and to a lower extent Costa Rica – have noticeable price differences between houses and apartments, which could potentially be attributed to composition effects as houses and apartments are not equally located in all places. The strength of observed correlation between house price and size also varies across countries (Figure 3). In South Africa, this correlation is weakest – among apartments, houses, and overall. This suggests that size may not be the primary driver of price, and that other attributes collected (e.g., location) may have more explanatory power.

Equally insightful are frequency distributions of smaller-sized apartments and houses within the data (Annex 4). Across countries, units smaller than 200 sqm are usually apartments. In Costa Rica and Pakistan, however, there are a significant number of smaller-sized houses, compared to the other emerging economies in our sample. In Pakistan, houses are dominated by 5-Marla[14] houses (equal to about 126 sqm), which are considered a typical house for a small family. 5-Marla houses are particularly prominent in Lahore, Rawalpindi, Islamabad, and Peshawar.

[14] The Marla is a traditional unit of area that is used in India, Pakistan, and Bangladesh, with one Marla being equal to 25.29 square meters.

Figure 1. Frequency Distribution of Property Type, by Country



Albania

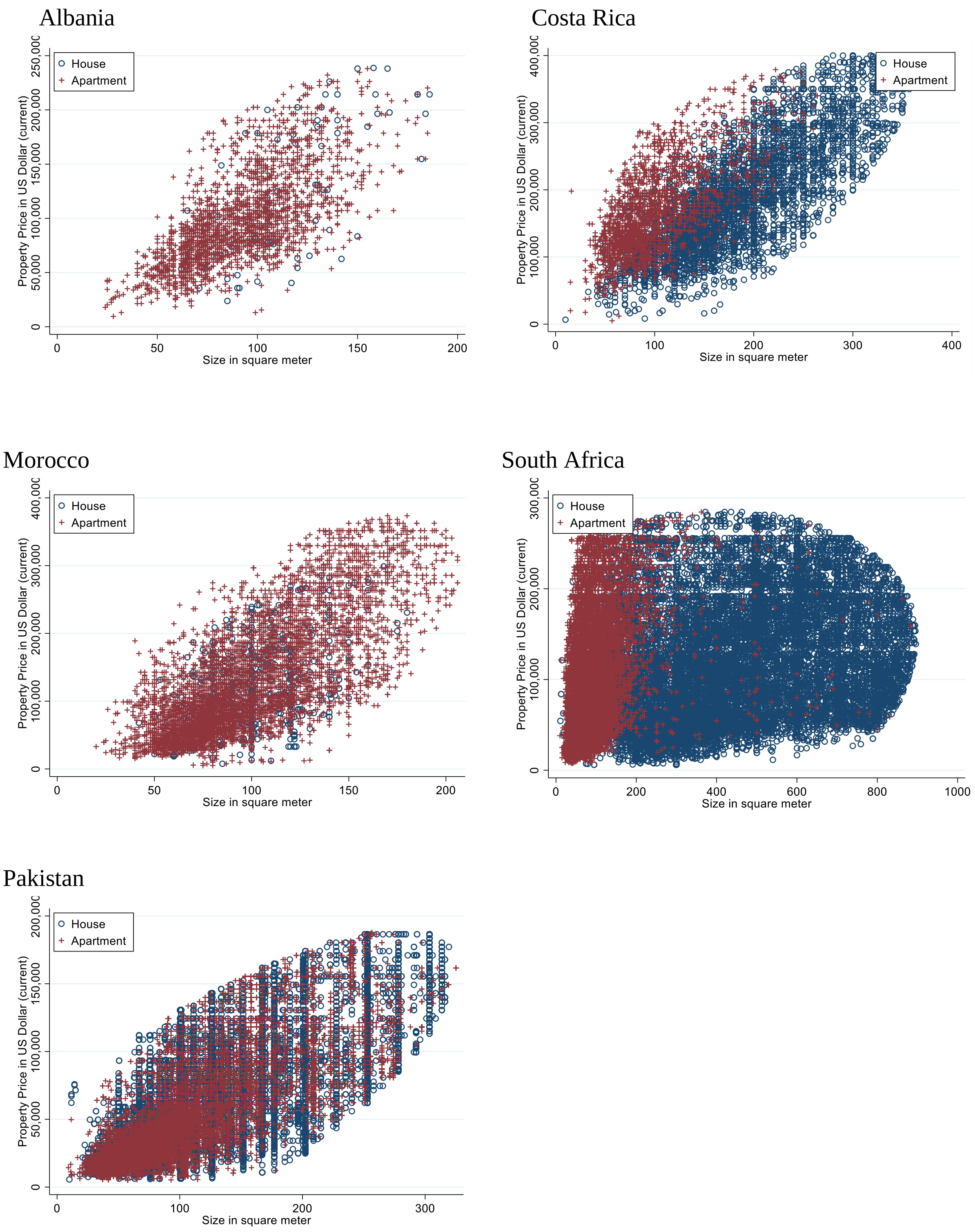
Costa Rica

Morocco

South Africa

Pakistan

Figure 2. Relationship of Price and Size, by Country



Albania

Costa Rica

Morocco

South Africa

Pakistan

4. Application: Estimation of the Basic House Price

This section demonstrates how the large volumes of property price data collected for select EMDEs can be used beyond descriptive statistics. To do so, we introduce the notion of the *Basic House Price (BHP)*, the price of a standard house that is defined identically across all markets. By fixing the type of house to be the constant, *BHP* aims to provide a data point on price that is independent of the distribution of type/quality of housing that varies widely across markets. The *BHP* is a key concept that allows for the comparison across and within EMDEs, assessing critical drivers of price and performance of housing markets at the lower end of the price spectrum.

We apply a machine learning technique, Random Forest (RF), to estimate the *BHP* from the collected web scaped data. Compared to Ordinary Least Squared (OLS) regression, Random Forest has consistently been found to perform better and provide more accurate price predictions (cf. Section 2.3). While the results are presented for five countries as way of application in the next section, the methodology introduced can be rolled out for all emerging economies.

4.1 Defining a Basic House

Housing costs reflect the value of the land, the price of the house, the age, condition, and location of the property, as well as the local market. Also, private property prices depend on macroeconomic as well as demographic conditions including migration, urbanization rates, population growth, income growth, a country's housing finance system including current interest rates and the availability of mortgage lending for all segments of the population. Other aspects relevant to house prices include challenges on the supply side such as a restrictive regulatory environment with lengthy permit granting processes, a shortage of labor and low mobility, as well as high construction costs.

In estimating the *Basic House Price*, we focus on formal housing built by private developers, through public-private partnerships between developers and governments, or by private individuals. We disregard projects that are purely government-sponsored, or housing that is self-built and probably not transacted in the housing market. Formal housing combines specializations in the housing value chain to deliver titled properties that can be pledged as collateral for a mortgage, that is structurally sound, and that complies with local planning standards and building codes (World Bank 2015). As formal housing often remains unaffordable to low-income HHs in many emerging economies, many families in these economies adhere to incremental self-building. Self-building is particularly common at the outskirts of larger cities or in smaller towns. A recent study in India, for instance, found that 62 percent of newly financed houses are self-built (Das et al. 2018). In self-built environments, the initial house serves as anchor for a multi-room home that accommodates multiple unrelated people or households (World Bank 2015). While these self-builders add to providing shelter to many families where the alternative is often homelessness, we disregard these houses for this project as self-built houses are often highly insecure in terms of tenure and do not necessarily comply with quality housing standards, building codes, or zoning regulations, and are not transacted in formal housing markets.[15] Also, we do not consider endogenous factors such as HH preferences over a set of amenities or locations, that might differ across HHs and countries. Given the limited information available and to maintain comparability across countries, we exclude these exogenous

[15] In Europe, in contrast, self-building is actively promoted as a means of addressing issues related to housing quality, affordability and sustainability (e.g., Bossuyt et al. 2018; Mullins & Moore 2018).

preferences from our model.[16] Finally, we are only concerned with home ownership and defer scraping of rental data to further research. The relationship between property prices and rental prices has been discussed in depth elsewhere (e.g., Campbell et al. 2009; Engsted & Pederson 2015; Gallin 2008).

We define a basic house as a

> *formally supplied 50 square meters (sqm) one-bedroom, one-bathroom apartment located in an urban area within a given country, assumed to provide basic municipal or on-site services including water, sanitation, road access, and an energy source.*

While we presented summary statistics for both houses and apartments, the *BHP* deliberately only includes apartments. We constrain the *BHP* to apartments for comparability purposes and to avoid potential distortions that can be attributed to the different reporting of size (plot size versus usable surface size) in houses. In addition, by reporting the *BHP* exclusively for apartments, we also account for the ongoing debate regarding the need to increase the housing density in emerging economies, particularly in cities that experience an influx of migration and growth, through densification of existing settlements or the building of multi-story, complex buildings. This is particularly relevant for Africa, where cities are 20 percent more fragmented compared to cities in Asia, more expensive and less accessible for most (Lall et al. 2017). In some African countries, the densification, which is not served by the market, takes place in the informal realm. In many countries, single family homes built on a plot of land are turned into mini-compounds where a main house is surrounded by 'backyard shacks' that are rented. This phenomenon of *backyarding* is particularly well documented for South Africa where *backyarding* increased from 1.1. million in 2011 to about 1.8 million in 2016 providing many families an informal way of overcoming the limitation of housing supply in urban areas (e.g., Brueckner et al. 2018).[17] Densification of houses has many beneficial effects, including a reduction of land use costs as well as cost of connecting to utility infrastructure and services, particularly in areas of accelerated urbanization (e.g., Kurvinen & Saari 2020). While we deliberately apply a narrow definition of the *BHP*, the presented methodology in this paper allows for easy transferability of other comparative units similar to *BHP* that might be more suitable for other researchers' focus.

4.2 Estimating the Basic House Price: Random Forest Estimation

To estimate the *BHP* for each country, we run the following predictive regression specification using house-level data that we obtain by web scraping online listings:

$\text{Price}_{,j,}= f(\text{Size}_{i,}\ \text{Type}_{i,}\ \text{Char}_{i,}\ \text{Location}_{i})$

where Price_i is the listed price of the property i; Size_i is the size in square meters; Char_i is a vector of characteristics of the property i to include the number of bedrooms and bathrooms; Location is a vector to denote the location on property i, and includes, where available, the municipality, county, and/or city.

The predictive framework above is estimated in its linear form using OLS and through non-parametric estimation of the RF model. When it comes to ML approaches to predicting house prices, there is an

[16] Recent research has applied deep learning models, especially convolutional neural networks (CNN), which allow the incorporation of the surroundings and other preferences into price estimation models (e.g., Law et al. 2019; Poursaeed et al. 2018).

[17] Backyarding is not just a phenomenon in emerging economies. In Los Angeles or Sydney, for instance, so-called *granny flats* or *casitas*, which are essentially small backyard houses, are encouraged by the city administration to combat the growing number of homeless people (e.g., Durst & Wegman 2017; Gurran et al. 2020).

expanding list of different approaches such as Random Forests, Quantile Regression, LASSO Regression, Adaptive Regression Splines, and Neural Nets (cf. Steurer et al. 2021), gradient boosting machine (GBM) or support vector machine (SVM) (e.g., Ho et al. 2021; Truong et al. 2020). Since previous research has demonstrated that Random Forest algorithms present the most accurate predictions, we decided for this non-parametric estimation technique and present other estimations for robustness checks (Mohd et al. 2019; Mullainathan & Spiess 2017; Pérez-Rave et al. 2019).

RF (Breiman, 2001) has recently gained popularity in property price predictions. RF models are based on classification and regression trees, which follow binary rule-based decisions that indicate how an input is related to its predictor variable (cf. Yoo et al. 2012). The RF is random in two ways: (1) each tree is based on a random subset of the observations, and (2) each split within each tree is created based on a random subset of variables (Grömping 2009: 311). In RF models, node splitting is not accomplished using all predictors as conventionally done in regression trees. Instead, RF node splitting is achieved using a random subset of predictors chosen at each node (e.g., Breiman 2001; Liaw & Wiener 2002). Hence, RF models are an ensemble tree-based learning algorithm that averages predictions over many individual trees using bootstrap aggregation (also known as bagging) (Breiman 2001).

Applied to the real estate sector, RF maps each vector of house characteristics to a predicted value. The prediction function takes the form of a tree that splits at every node given the value of a particular housing characteristic (e.g., sqm; number of rooms) (Mullainathan & Spiess 2017). Given its very flexible functional form, RF is suited well for out-of-sample predictions and for varied structures of data. Unlike other econometric estimation techniques, RF models do not require training data to be normally distributed, which particularly for property price research in EMDEs is beneficial as data might be heavily skewed.

While relatively new to property price estimations, RF models have a variety of advantages over traditional estimation techniques, particularly in EMDEs. First, compared to other price estimation models, RF models perform stronger than other algorithms, offering more precise price estimations (cf. Section 2.3). Second, housing markets in EMDEs often have a series of sub-markets either clustered around housing size, type of housing, or income group. Traditional estimations like hedonic price models would often fail to capture these sub-markets. Hence, if the data set sufficiently covers the characteristics of the property, the RF model is expected to replicate the complex structure of the property price determination process more sensitively (cf. Hong 2020: 142). Third, RF models do not require a detailed model and are hence more suitable for EMDEs with potentially more skewed distributions. Finally, while hedonic price models have been more geared towards inference, RF models focus more on prediction (Yoo et al. 2012).

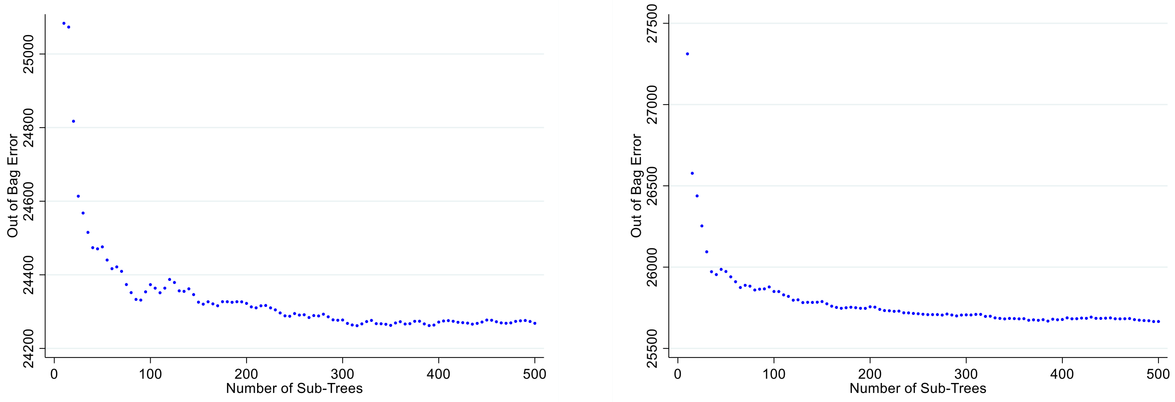
4.3 Parameter Optimization

The model-training process is started by randomly splitting the data set into training and testing data for each country ensuring a random sort order. We split each country's data set into two subsets: 50 percent of the data are used for training, and 50 percent of the data are used for testing (validation) (cf. Schonlau & Zou 2020). The 50-50 split is the most common split in RF applications. Results on alternate splits are also tested and presented in Annex 5. More in-depth discussions on the effect of alternate splitting options are offered elsewhere (cf. Biau 2012; Ishwaran 2014).

Having decided on splitting the data set into training and testing data, we tune the hyperparameters to determine the model with the highest testing accuracy, focusing on the number of sub-trees and the number of variables randomly investigated at each split. The benefit of RF is that there are few hyperparameters with the potential to strongly influence the model's performance (cf. Hong 2020). RF does not require an external cross-validation procedure to estimate the model's accuracy. Model selection and parameter tuning are driven by parameters that would produce the lowest out-of-bag (OOB) errors.[18]

First, we fix the number of sub-trees (number of iterations). As RF OOB error rates converge after the number of iterations gets large enough, we set the iterations to 500 for all models instead of tuning the number of observations for each country's data set individually (cf. Breiman 2001; Schonlau & Zou 2020). While some scholars applying RF spend a fair amount of time in tuning to the most optimal number of sub-trees, recent research has shown that increasing the number of trees does not harm the model and the biggest performance gain is achieved within the first 250 trees (Probst & Boulesteix 2018). To check for the robustness of this assumption for our data, we iteratively run the model for two countries testing incrementally how increasing iterations from 10 to 500 alters the OOB error rate. As error rates stabilize with increasing iterations for both countries, we chose the highest number of sub-trees (500) for all our models (Figure 3).

Figure 3. Out-of-Bag Error Rate for Varying Number of Iterations



Albania **Morocco**

Second, we select the number of variables to randomly investigate at each split – the depth of the decision trees. RF models applied to property price estimations in developed markets often have many property attributes to choose from (e.g., square meters, bedroom, bathroom, garage, age of property, location, distance to markets etc.). In these scenarios, to select the best RF model, authors often remove lesser important property attributes one at a time to estimate the relative performance of the model (e.g., Hong 2020) or "only" use the ten most important predicting variables in the final model (e.g., Čeh et al. 2018).

[18] The error of the Random Forest is approximated by the OOB error during the training process. Each tree is built on a different bootstrap sample which, by random chance, leaves out about one-third of the observations. These left-out observations for a given tree are referred to as the OOB sample. Finding parameters that would produce low OOB error is often a key consideration in model selection and parameter tuning (cf. Schonlau & Zou 2020: 6).

Selecting the number of attributes where the OOB-error rate is lowest is another common decision factor in RF model selection (Schonlau & You 2020).

Since the number of property attributes is rather limited in most EMDEs that we cover, we include all available attributes to the RF model. For most of the five EMDEs presented in this paper, this includes at least the type of the property (apartment or house), size of the property, location (city, region, or district – depending on availability), number of bedrooms, and number of bathrooms. Costa Rica, unfortunately, does not provide the number of bedrooms and bathrooms, and hence only Size, Type of Property, and City are included as predictor variables. The exact variables used for each country are summarized in Table 2.

Table 2. Variables used in RF Model

Country	Variables used in RF Model
Albania	Size, Type of Property, Number of Bedroom, Number of Bathroom, County, City
Costa Rica	Size, Type of Property, City
Morocco	Size, Type of Property, Number of Bedroom, Number of Bathroom, City
Pakistan	Size, Type of Property, Number of Bedroom, Number of Bathroom, City
South Africa	Size, Type of Property, Number of Bedroom, Number of Bathroom, Province, Municipality, City

Since all property attributes that we are using for property price estimations have consistently been found to be relevant for price predictions (cf. Section 2.2) and since the number of property attributes is overall limited, we abstain from successively identifying the optimal number of features in the RF model and include all available attributes in our final model.[19]

5. Results: Private Property Prices in Five EMDEs

In the following section, we discuss the results of private property prices in five emerging economies across different regions: Albania, Costa Rica, Morocco, Pakistan, and South Africa. We selected these economies as a way of demonstrating how a big data approach can be applied to notoriously data-scarce environments such as EMDEs.

5.1 Basic House Prices in Five Economies and Their Largest Cities

Having provided some overview on the availability of houses and apartments in the market and having discussed some descriptive statistics on price and size of all available private properties within the available data, we now present the estimation of the *BHP*, which solely focuses on apartments. Estimating the property price of a Basic House as defined in Section 4.1, Table 3 summarizes the results of the estimation

[19] To check for the robustness of this approach, we pooled all countries' data into a global data set and trained the RF model on the overall data applying the same parameter optimization. Allowing for transfer learning of the model across countries, we then provide country-specific estimates derived from the global data set. Results of this approach are broadly in line for all countries except Albania, where we attribute the deviation of results to the comparably smaller number of observations compared to other countries. Deviation in number of observations across countries, as we are observing in our model, may contribute to bias when a global RF model is applied (cf. Annex 6).

model. To compare the performance, we use the same explanatory variables available for every country across models (as outlined in Table 2). All results are robust to more rigorous removal of potential duplicates in the data (Annex 2) as well as the application of alternate splits (Annex 3).

Local property markets have their own characteristics featuring from the market itself and the products offered. The national averages in Table 3 mask the differences within the country. Particularly within capitals or the biggest business city, house prices are expected to be more expensive than in less urbanized areas. To capture the different price dynamics, we provide price estimations for the *BHP* for the countries' largest cities in Table 3.

Table 3. Basic House Prices, by City

	Random Forest-Based Prediction						OLS-Based Prediction					
Country	Basic House Price *Apartment (Current USD)*	Basic House Price *Apartment (Current PPP$)*	Ratio predicted and obs. price *(median)*	Local Ratio predicted and obs. price *(median)*	MAPE	L_ MAPE	Basic House Price *Apartment (Current USD)*	Basic House Price *Apartment (Current PPP$)*	Ratio predicted and obs. price *(median*	Local Ratio predicted and obs. price *(median)*	MAPE	L_ MAPE
Albania												
National	71,205	189,200	1.01	1.01	0.28	0.19	55,167	146,592	1.02	1.07	0.28	0.26
Durres	45,422	120,697	1.05	0.94	0.35	0.20	41,359	109,900	1.16	1.05	0.41	0.26
Sarande	73,794	196,088	1.01	.	0.38	.	49,601	131,801	1.04	.	0.15	.
Tirana	74,337	197,531	1.00	1.02	0.26	0.19	56,649	150,529	1.00	1.00	0.26	0.25
Vlore	54,856	145,765	1.27	0.85	0.39	0.16	53,627	142,499	1.38	0.90	0.46	0.17
Costa Rica												
National	113,065	194,476	0.98	0.97	0.27	0.40	115,139	198,044	0.99	1.06	0.30	0.43
Alajuela	110,284	189,694	1.23	.	0.38	.	107,774	185,376	1.32	.	0.49	.
Heredia	104,378	179,535	1.08	1.16	0.22	0.14	111,794	192,290	1.11	1.19	0.22	0.22
Puntarenas	130,370	224,242	1.00	.	0.01	.	115,813	199,204	0.87	.	0.13	.
San José	106,457	183,111	0.85	0.82	0.23	0.20	116,961	201,179	0.90	0.91	0.22	0.18
Morocco												
National	53,282	129,554	1.00	1.04	0.30	0.29	47,201	114,769	1.01	0.96	0.42	0.41
Agadir	40,081	97,456	1.07	1.12	0.37	0.33	53,670	130,570	1.27	1.61	0.54	0.66
Casablanca	79,739	193,886	0.93	1.07	0.26	0.30	64,356	156,482	0.86	0.60	0.24	0.38
Fez	42,155	102,500	1.05	.	0.31	.	41,762	101,544	2.05	.	1.18	.
Tangier	40,828	99,272	0.99	1.04	0.31	0.28	28,052	68,209	0.93	0.73	0.31	0.28
Marrakesh	48,856	118,793	1.00	1.09	0.26	0.25	43,852	106,626	0.94	0.82	0.26	0.28
Pakistan												
National	21,849	91,269	0.99	1.16	0.31	0.30	24,360	101,758	1.00	1.27	0.44	0.52
Islamabad	21,711	90,692	0.99	0.94	0.30	0.29	24,607	102,789	1.05	1.01	0.38	0.33
Karachi	21,646	90,422	0.98	1.18	0.34	0.33	24,380	101,842	0.91	1.35	0.41	0.48
Lahore	26,651	111,328	1.00	1.15	0.28	0.32	24,153	100,895	1.05	1.06	0.50	0.31
Rawalpindi	25,837	107,927	1.02	1.10	0.25	0.24	23,519	98,244	1.06	1.01	0.34	0.22
South Africa												
National	63,745	137,501	1.02	1.07	0.30	0.30	85,769	185,006	1.15	1.27	0.55	0.56
Cape Town	106,536	229,801	0.89	1.02	0.36	0.33	132,455	285,709	0.75	0.96	0.42	0.37
Durban	43,917	94,731	1.03	1.05	0.32	0.28	81,276	175,314	1.28	1.49	0.53	0.63
Johannesburg	60,826	131,204	1.05	1.18	0.64	0.76	86,642	186,889	1.21	1.50	0.91	0.85

Note: MAPE refers to mean absolute percentage error; l_MAPE refers to the mean percentage error based on predictions accuracy of apartments sized between 50 and 60 square meters; local ratio refers to the ratio between observed and predicted values for apartments sized between 50 and 60 square meters. Exchange rates are based on 2019 conversions.

The estimations show that among the five countries, at the national level and in US$ terms, Pakistan has the cheapest *BHP* (US$ 21,849), followed by Morocco (US$ 53,282), South Africa (US$ 63,745), Albania (US$ 71,205), and Costa Rica (US$ 113,064). In PPP$, however, the price levels for a basic apartment between the five countries is more equal. Visualization of the raw data in Figures 1 and 2 offer some intuition behind the cross-country differences. Comparing Pakistan and Costa Rica at opposite ends of the spectrum, we observe the availability of apartments in Pakistan concentrated at the lower end of the price spectrum, while a much more even distribution in Costa Rica (Figure 1). Moreover, while the price-size relationship (Figure 2) clearly points to apartments being listed cheaper than similar size houses in Pakistan, the opposite is true in Costa Rica. As the BHP represents the typical estimated market price for a standard 50 sqm apartment, the prevalence of more luxurious/expensive apartments in Costa Rica is expected to drive up the benchmark price.

Within countries, there are also significant differences across regions and cities. In South Africa, for instance, price differences of the BHP between cities are stark. House prices in Cape Town are at the higher income spectrum, where even a basic house is priced at US$ 106,536. Cape Town is one of the most popular tourist destinations in Africa and its property market is known to be tailored to more affluent retirees and foreign property buyers. In Morocco, Casablanca is the most expensive city followed by Marrakesh. In addition to consistently high prevalence of European buyers, many wealthy Moroccans families live in the suburbs of these cities including Palmeraie in Marrakesh and Bouskoura in Casablanca where prices usually start around US$ 700,000. In Pakistan, while nationally at the lowest end of the price spectrum of the five economies covered in this study, there are also significant intra-country differences. Being a fast-growing emerging economy, the capital, Islamabad, is a thriving real estate market. While there are many houses at the lower end of the price spectrum, with the cheapest house advertised at roughly US$ 9,000, property prices in Islamabad can go as high as US$ 6 million.

5.2 Cross-Validation

The predicted *BHP* of the apartments obtained by both models were compared with the observed apartment prices in order to determine the predictive power of the different models. One standard measure often used in price estimation models is the quotient between the predicted price and the observed price for the property. The acceptable median ratio between predicted and observed price is 0.9-1.1 (cf. International Association of Assessing Officers 2014; Čeh et al. 2018).

Additionally, we evaluate the performance of the different models with the mean absolute percentage error (MAPE), which measures the average percentage deviation of predicted prices from actual property prices expressed as

$$\text{MAPE}= \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{\mathbf{p}}-\boldsymbol{pi}}{\boldsymbol{pi}} \right|,$$

where $\hat{p}$ is the predicted property price and p_i the actual property price of property i. To understand differences in predictions for the BHP, we estimate a localized MAPE for apartments between 50 and 60 square meters. Comparing the predictive performances of the models based on the performance measures (ratio of predicted vs. actual value, MAPE, localized MAPE), we obtain more precise results in the RF models for all countries. All RF models are within the suggested range of the predicted/observed price ratio of 0.9-1.1, also for the estimations where all potential duplicates are rigorously removed (Annex 2) and where different splits are applied (Annex 5). In the main OLS estimation, South Africa exceeds the acceptable median ratio between predicted and observed price range. The MAPEs indicates that the percent deviation of the RF prediction from the actual property price ranges between 27 percent for Costa Rica and 31 percent for Pakistan. Across the board, while MAPE is relatively high in both RF-based and OLS-based predictions, the RF estimation consistently performs better than the OLS prediction. While the quality and quantity of information used in both estimation techniques were identical, since the predictor in the RF model explores the hierarchical structure of features, it can more sensitively track the possibility that the effect of each attribute on price varies by context (Hong 2020).

The limited coverage of observable property features within EMDEs could be one explanation for the acceptable accuracy of the estimation. In addition, while most studies applying the RF model focus on a very narrow housing market (e.g., Čeh et al. 2018; Levantesi & Piscopo 2020) our data expands to the entire housing market of the five emerging economies covering heterogeneous properties with varying amenities including interior decorations, building age, or other features that are not captured in the model as these property features were not consistently available on listing websites. Equally, we are jointly estimating property prices for a large swathe of locations within an economy – beyond just different neighborhoods within a city but aggregating both rural and urban areas. This deviates from the use of RF models in the literature to predict property prices for a well-defined narrow set of locations, typically a city or a province/state. The limited property features and available explanatory variables pose limitations to the use of the model to predict individual property prices across the spectrum. However, the purpose of the prediction in our case is to arrive at a typical price for a standard property that can be compared across countries and contexts. The RF and OLS-based approach, notwithstanding the relatively high MAPE, can be considered improvements over the alternative of only considering one-dimensional summary statistics of price.

6. Conclusion

Given the difficulty of obtaining reliable private property price data in emerging economies, most analyses of house prices or affordability assessments are constrained to developed economies and limited in scope. Most cross-country analyses that assess trends in house prices are based on available indices, which often aggregate to the national level, masking important within-country dynamics and regional differences. To overcome this flaw and provide more in-depth insights into housing markets in emerging economies, we demonstrate how to collect a large range of localized data through web scraping of property listing websites. Further, to compare property prices across countries, we introduce the concept of *the Basic House Price* – which constitutes the average price of a basic one-bedroom apartment of 50 square meters in an urban area – that allows for comparability across countries.

By way of demonstrating the methodology and data processing for five EMDEs, we show that web scraping offers a cost-effective way to obtain a large amount of price data for countries where official data is absent and where alternate data sources on prices are not available. The main constraint to this approach remains the unorganized structure of listing websites and the limited information available on property features and attributes. This approach will only improve over time as the capabilities of listing websites improve and become the preferred method of listing. There is also room to improve the web scraping on several fronts. For instance, image recognition software can extract information that is not supplied systematically in listing websites and could improve model precision. In addition, addresses could be geotagged to incorporate crucial details about location and to differentiate within-city variation.

The paper aims to outline one efficient way to address the wide data gap on property prices in emerging economies. In addition, the paper outlined how, once collected, these data could be used to estimate the price of a standard house consistently. With a consistent methodology proposed in our paper, the *BHP* can then be applied in several avenues for further research.

First, available data and analysis can feed into further research on determinants of house prices and drivers of changes through time in emerging economies. While determinants of house prices are a well-researched subject in the literature, gaining increasing attention post-2008/2009, analyses mainly rely on data from developed economies. If available at all, price estimations in emerging economies are primarily available for very localized markets. To what extent these findings extend to a larger sample of emerging economies may be an area of research triggered by the data proposed in this paper. In addition, research areas more relevant for emerging economies, such as those related to empirically assessing inefficiencies in the housing value chain, would be possible with the data and analysis proposed by the paper. From an affordability perspective, this paper provides an important variable that may be combined with other data sources, for example, households' disposable income. Bringing these various elements together in the analysis of the country's housing market affordability is fundamental for more fully understanding housing needs and challenges faced by households in emerging economies.

References

Ardila, D., Ahmed, A., & Sornette, D. (2021). Comparing Ask and Transaction Prices in the Swiss Housing Market. *Quantitative Finance and Economics* 5(1) 67-93.

Alter, A., & Mahoney, E. M. (2021). Local House-price Vulnerability: Evidence from the U.S. and Canada. *Journal of Housing Economics* 54, 1-17.

Anenberg, E. & Laufer, S. (2017). A More Timely House price Index. *The Review of Economics and* Statistics 99(4), 722-734.

Anderson, D. E. (2000). Hypothesis Testing in Hedonic Price Estimation. On the Selection of Independent Variables. *The Annals of Regional Science* 34(2), 293-304.

Anundsen, A. K., Gerdrup, K. & Hansen, F. (2016). Bubbles and Crises: The Role for House Prices and Credit. *Journal of Applied Econometrics* 31(7), 1291-1311.

Angelo, H., & Vormann, B. (2018). Long Waves of Urban Reform: Putting the Smart City in its Place. *City* 22(5–6), 782-800.

Antipov, E.A. & Pokryshevskaya, E.B. (2012). Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-based Approach for Model Diagnostics. *Expert Systems with Applications* 39(2), 1772-1778.

Anenberg, E. & Laufer, S. (2017). A More Timely House Price Index. *The Review of Economics and Statistics* 99(4),722‑734.

Arnott, R. (2009). Housing Policy in Developing Countries: The Importance of the Informal Sector, in Spence, M., P.C. Annex & R.M. Buckley (eds.). *Urbanization and Growth.* Washington, DC: Commission on Growth and Development, pp. 167‑97.

Athey, S. (2018). The Impact of Machine Learning on Economics. In A. K. Agrawal, J. Gans, & A. Goldfarb (eds.). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, pp. 507-547.

Balcilar, M., Gupta, R., & Shah, Z. B. (2011). An In-sample and Out-of-sample Empirical Investigation of the Nonlinearity in House Prices of South Africa. *Economic Modelling* 28(3), 891-899.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research* 13, 1063–1095.

Boeing, G., Besbris, M., Schachter, A., & Kuk, J. (2021) Housing Search in the Age of Big Data: Smarter Cities or the Same Old Blind Spots*? Housing Policy Debate*, 31(1), 112-126.

Borba, J. O. & Dentinho, T. P. (2016). Evaluation of Urban Scenarios Using Bid-rents of Spatial Interaction Models as Hedonic Price Estimators: An Application to the Terceira Island, Azores. *The Annals of Regional Science* 56(3), 671-685.

Borde, S., Rane, A., Shende, G., & Shetty, S. (2017). Real Estate Investment Advising Using Machine Learning. *International Research Journal of Engineering and Technology* 4(3), 1821-1825.

Borio, C., Kennedy, N. & Prowse, S. (1994). Exploring Aggregate Asset Price Fluctuations across Countries: Measurement, Determinants and Monetary Policy Implications. BIS Economic Papers, Basle, Switzerland: BIS.

Bossuyt, D., Salet, W. & Majoor, S. (2018). Commissioning as Cornerstone of self-build Housing. Assessing the Constraints and Opportunities of self-build in The Netherlands. *Land Use Policy* 77, 524-533.

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5-32.

Bricongne, J-C., Turrini, A., & Pontuch, P. (2019). Assessing House Prices: Insights from 'Houselev', A Dataset of Price Level Estimates. *European Economy Discussion Papers 101*.

Bricongne, J-C., Meunier, B., & Pouget, S. (2021). Web Scraping Housing Prices in Real-time: the Covid-19 Crisis in the UK, *Banque de France Working Paper,* No. 827.

Brueckner, J. K., Rabe, C., & Harris, S. (2018). Backyarding: Theory and Evidence for South Africa. *Policy Research Working Paper No. 8636*. Washington, D.C.: World Bank.

Campbell, D., Morris, D., Gallin, J., & Martin, R. (2009). What Moves Housing Markets: A Variance Decomposition of the Rent-Price Ratio. *Journal of Urban Economics* 66, 90-102.

Campbell, J. Y. & Cocco, J. F. (2007). How do House Prices affect Consumption? Evidence from Micro Data. *Journal of Monetary Economics* 54 (3), 591–621.

Can, A. (1992). Specification and Estimation of Hedonic Housing Price Models. *Regional Science and Urban Economics* 22(3), 453-474.

Catte, P., Price, R. W. R., Girouard, N., & André, C. (2004). Housing Markets, Wealth and the Business Cycle, *OECD Economics Department Working Paper*, No. 394. Paris: OECD.

Ceccato, V., & Wilhelmsson, M. (2020). Do Crime Hot Spots Affect Housing Prices? *Nordic Journal of Criminology* 21(1), 84-102.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *International Journal of Geo-Information*, 7(5).

Claessens, S. & J. Schantz (2019). Regional House Price Differences: Drivers and Risks. In Nijskens, R., Lohuis, M., Hilbers, P., Heeringa, W. (Eds.) *Hot Property. The Housing Market in Major Cities*. Cham, CH: Springer, pp. 39-49.

Das, C., A. Karamchandani, & Thuard, J. (2018). *State of the Low-Income Housing Finance Market*. Boston: FSG.

Davis, M. A. & Heathcote, J. (2005). Housing and the Business Cycle. *International Economic Review* 46(3), 751-784.

Deghi, A., Katagiri, M., Shahid, S., Valckx, N. (2020). Predicting Downside Risks to House Prices and Macro-Financial Stability. *International Monetary Fund* WP/20/11.

Del Negro, M. & Otrok, C. (2007). 99 Luftballons: Monetary Policy and the House Price Boom across U.S. States. *Journal of Monetary Policy* 54(7), 1962-1985.

Durst, N. J. & Wegmann, J. (2017). Informal Housing in the United States. *International Journal of Urban and Regional Research* 41(2), 282-297.

Drehmann M. & Juselius M. (2014). Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements. *International Journal of Forecasting* 30, 759-780.

Engsted, T. & Pedersen, T.Q. (2015). Predicting Returns and Rent Growth in the Housing Market Using the Rent-Price Ratio: Evidence from the OECD Countries. *Journal of International Money and Finance* 53, 257-275.

European Union, International Labour Organization, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations Economic Commission for Europe, The World Bank (2013). *Handbook on Residential Property Prices Indices (RPPIs)*. Luxembourg European Union.

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies* 43(12), 2301-2315.

Fletcher, M., Gallimore, P., & Mangan, J. (2000). Heteroscedasticity in Hedonic House Price Models. *Journal of Property Research* 17(2), 93-108.

Follain, J. R., Jr. & Malpezzi, S. (1981). Are Occupants Accurate Appraisers? *Review of Public Data Use* 9 (1), 47-55.

Gardner, D. & Pienaar, J. (2019). Benchmarking Housing Construction Costs in Africa. Centre for Affordable Housing Africa. http://housingfinanceafrica.org/app/uploads/Benchmarking-Housing-Construction-Costs-Across-Africa-FINAL-19-May-2019.pdf

Gallin, J. (2008). The long-run Relationship Between House Prices and Rents. *Real Estate Economics* 36, 635-658.

Gauder, M., Houssard, C., & Orsmond, D. (2014). Foreign Investment in Residential Real Estate. *Reserve Bank of Australia Bulletin*. https://www.rba.gov.au/publications/bulletin/2014/jun/pdf/bu-0614-2.pdf

Gao, G., Bao, Z., Cao, J., Quin, A.K., Sellis, T. & Wu, Z. (2019). Location Centered House Price Prediction: A Multi-Task Learning Approach. arXiv arXiv:1901.01774.

Garrod, G. D. & Willis, K. G. (1992). Valuing Goods' Characteristics: An Application of the Hedonic Price Method to Environmental Attributes. *Journal of Environmental Management* 34(1), 59-76.

Girouard, N., Kennedy, M., van den Noord, P., & André, C. (2006), Recent House Price Developments: The Role of Fundamentals, OECD Economics Department Working Papers No. 475.

Glaeser, E L., & Ward, B. A. (2009). The Causes and Consequences of Land Use Regulation: Evidence from Greater Boston. *Journal of Urban Economics* 65(3), 265-278.

Gnagey, M. & Tans, R. (2018). Property-Price Determinants in Indonesia. *Bulletin of Indonesian Economic Studies* 54(1), 61-84.

Goodhart, C. & Hofmann, B. (2008). House Prices, Money, Credit, and the Macroeconomy. *Oxford Review of Economic Policy* 24 (1):180-205.

Goodman, J. L. & Ittner, J. B. (1992). The Accuracy of Home Owners' Estimates of House Value. *Journal of Housing Economics* 2 (4), 339-57.

Gröbel, S. & Thomschke, L. (2018). Hedonic Pricing and the Spatial Structure of Housing Data - An Application to Berlin. *Journal of Property Research* 35(3), 185-208.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *American Statistical Association* 63(4), 308-319.

Gurran, N., Maalsen, S., & Shrestha, P. (2020). Is Informal Housing an Affordability Solution for Expensive Cities? Evidence from Sydney, Australia. International Journal of Housing Policy

Gyourko, J., Mayer, C. & Sinai, T. (2013). Superstar Cities. *American Economic Journal* 5(4), 167-199.

Haurin, D. R., Haurin, J. L., Nadauld, T. & Sanders, A. (2010). List Prices, Sale Prices and Marketing Time: An Application to U.S. Housing Markets, *Real Estate Economics* 38 (4), 659–85.

He, S., Wang, D., Webster, C., & Chau, K. (2019). Property Rights with Price Tags? Pricing Uncertainties in the Production, Transaction and Consumption of China's Small Property Right Housing. *Land Use Policy* 81, 424-434.

Ho, W. K. O., Tang, B.-S., Wong, S. W. (2021). Predicting Property Prices with Machine Learning Algorithms. *Journal of Property Research* 38(1),48-70.

Hong, J., Choi, H., Kim, W. (2020). A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *International Journal of Strategic Property Management* 24(3), 140-152.

Horwitz, J. L. (1992). The Role of the List Price in Housing Markets: Theory and Econometric Model, *Journal of Applied Econometrics* 7, 115-129.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring Housing Rental Prices Based on Social Media: An Integrated Approach of Machine-Learning Algorithms and Hedonic Modeling to Inform Equitable Housing Policies. Land Use Policy 82, 657‑673.

ICP (2011). A New Approach to International Construction Price Comparison. Available at: http://siteresources.worldbank.org/ICPINT/Resources/270056-1255977254560/6483625-1273849421891/110622_ICP-OM_Construction.pdf

Igan D. & Loungani, P. (2012). *Global Housing Cycles*, IMF Working Paper, No. 12/217. Washington D.C.: International Monetary Fund.

International Association of Assessing Officers (2014). Guidance on International Mass Appraisal and Related Tax Policy. Available at: http://www.iaao.org/media/Standards/International_Guidance.pdf

International Monetary Fund. (2018). House Price Synchronization: What Role for Financial Factors? In IMF Global Financial Stability Report, April 2018. *A Bumpy Road Ahead* (pp. 93-133).

International Monetary Fund. (2008). *World Economic Outlook April 2008. Housing and the Business Cycle*. Washington, D.C.: International Monetary Fund.

Ishwaran, H. (2014). The Effect of Splitting on Random Forests. *Machine Learning* 99, 75-118.

Jordà, O., Schularick, M., & Taylor, A.M. (2016). The Great Mortgaging: Housing Finance, Crises and Business Cycles. *Economic Policy* 31(85), 107-152.

Jordà, O., Schularick, M., & Taylor, A.M. (2015). Betting the House. *Journal of International Economics* 96(S2), 2-18.

Keskin, B. & Watkins, C. (2017). Defining Spatial Housing Submarkets: Exploring the Case for Expert Delineated Boundaries. *Urban Studies* 54(6), 1446-1462.

Kim, A. M. (2007). North versus South: The Impact of Social Norms in the Market Pricing of Private Property Rights in Vietnam. *World Development* 35(12), 2079-2095.

Knoll, K., Schularick, M. & Steger, T. (2017). No Price Like Home: Global House Prices, 1870-2012. *American Economic Review* 107(2), 331-353.

Krol, A. (2013). Application of Hedonic Methods in Modelling Real Estate Prices in Poland. *Data Science, Learning by Latent Structures, and Knowledge Discovery*, 501-511.

Kurvinen, A. & Saari, A. (2020). Urban Housing Density and Infrastructure Costs. *Sustainability* 12(2).

Law, S., Paige, B., & Russell, C. (2019). Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10(5), 1-19.

Lall, S. V., Henderson, J. V., & Venables A. J. (2017). *Africa's Cities: Opening Doors to the World*. Washington, D.C.: The World Bank.

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy* 74 (2), 132-157.

Levantesi, S. & Piscopo, G. (2020). The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. *Risks* 8(4),1-17.

Li, M., Zhang, G., Chen, Y., Zhou, C. (2019). Evaluation of Residential Housing Prices on the Internet: Data Pitfalls. *Complexity* 1-15.

Liaw, A. & Wiener, M. (2002). Classification and Regression by Random Forest. *R News* 2(3), 18-22.

Libertun de Duren, N. R. (2018). Why There? Developers' Rationale for Building Social Housing in the Urban Periphery in Latin America. *Cities* 72, 411-420.

Luüs, C. (2005). The Absa Residential Property Market Database for South Africa: Key Data Trends and Implication. *BIS Papers 21*. Available at: https://www.bis.org/publ/bppdf/bispap21l.pdf .

Lyons, R. C, (2019). Can List Prices Accurately Capture Housing Price Trends? Insights from Extreme Market Conditions. Finance Research Letters 30, 228-323.

Mack, A. & Martínez-García, E. (2011). A Cross-Country Quarterly Database of Real House Prices: A Methodological Note. *Federal Reserve Bank of Dallas Globalization and Monetary Policy Institute Working Paper No. 99* https://www.dallasfed.org/~/media/documents/institute/wpapers/2011/0099.pdf

McGreal, S. & Taltavull de La Paz, P. (2013). Implicit House Prices: Variation over Time and Space in Spain, *Urban Studies* 50 (10), 2024-43.

McKinsey Global Institute (2014). A Blueprint for Global Affordable Housing Challenge. McKinsey Global Institute.

Mian, A., Sufi, A., & Verner, E., 2017. Household Debt and Business Cycles Worldwide. *The Quarterly Journal of Economics* 132(4), 1755-1817.

Montero, J.-M., Mínguez, R., & Fernández-Avilés, G. (2018). Housing Price Prediction: Parametric Versus Semi-parametric Spatial Hedonic Models. *Journal of Geographical Systems* 20, 27-55.

Mohd, T., Masrom, S., & Johari, N. (2019). Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia. *International Journal of Recent Technology and Engineering* 8(2S11), 542–546.

Mullainathan, S. & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2), 87-106.

Mullins, D. & Moore, T. (2018). Self-organised and Civil Society Participation in Housing Provision. *International Journal of Housing Policy* 18, 1-14.

Nairametrics (2018). Deal: ToLet.com.ng acquires Jumia House Nigeria, now Property Pro. Available at : https://nairametrics.com/2017/11/12/deal-tolet-com-ng-acquires-jumia-house-nigeria-now-property-pro/

Oladunni, T. & Sharma, S. (2016). *Hedonic Housing Theory. A Machine Learning Investigation*. International Conference on Machine Learning and Applications (pp. 522–527). Anaheim, United States, 18-20 December 2016.

Owusu-Manu, D., Edwards, D. J., Donkor-Hyiaman, K. A., Asiedu, R. O., Hosseini, M. R., Obiri-Yeboah, E. (2019). Housing Attributes and Relative House Prices in Ghana. *International Journal of Building Pathology and Adaptation* 37(5), 733-746.

Pérez-Rave, J., Correa-Morales, J. C. & González-Echavarría, F. (2019). A Machine Learning Approach to Big Data Regression Analysis of Real Estate Prices for Inferential and Predictive Purposes. *Journal of Property Research* 36(1), 59-96.

Pfeifer, N. & Steurer, M. (2020) Early real Estate Indicators During the COVDI-19 Crisis: A Tale of Two Cities. Graz Economic Papers, http://www100.uni-graz.at/vwlwww/forschung/RePEc/wpaper/2020-17.pdf

Philiponnet, N. & A. Turini (2017). Assessing House Price Developments in the EU. Discussion Paper 048 https://ec.europa.eu/info/sites/info/files/dp048_en.pdf

Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based Real Estate Price Estimation. *Machine Vision and Applications* 29 (4), 667–676.

Probst, P. & Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research 18, 1-18.

Rae, A. (2015). Online Housing Search and the Geography of Submarkets. *Housing Studies* 30(3), 453-472.

Reuter (2021). Coronavirus Wave flattens Indian Housing Market Views: Reuters poll. Available at: https://www.reuters.com/article/us-india-property-poll/coronavirus-wave-flattens-indian-housing-market-views-reuters-poll-idUSKCN2D20A4.

Rodriguez, M., & Sirmans, C. F. (1994). Quantifying the Value of a View in Single-Family Housing Markets. *Appraisal Journal* 62, 600-603.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82 (1), 34–55.

Sadayuki, T. (2018). Measuring the Spatial Effect of Multiple Sites: An Application to Housing Rent and Public Transportation in Tokyo, Japan. *Regional Science and Urban Economics* 70, 155-173.

Schonlau, M. & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal* 20(1), 3-29.

Steurer, M., Hilll, R. J., Pfeifer, N. (2021). Metrics for Evaluating the Performance of Machine Learning Based Automated Valuation Models. Journal of Property Research 38(2), 99-129.

ten Bosch, O. & Windmeijer, D. (2014). On the Use of Internet Robots for Official Statistics, UNECE MSIS conference, Dublin, Ireland 2014.

Truong, Q., Nguyen, M., Dang, H. & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science* 174, 433-442.

UN-Habitat. (2007). *Principles and Recommendations for Population and Housing Censuses* (revision 2). New York: United Nations.

Uwayezu, E. & de Vries, W. T. (2020) Access to Affordable Houses for the Low-Income Urban Dwellers in Kigali: Analysis Based on Sale Prices. *Land* 9(3), 2-32.

Veradi and McCathie (2012) The S-estimator of Multivariate Location and Scatter in Stata. *The Stata Journal*, 12(2), 299-307.

Wang, C. C. & Wu, H. (2018). A New Machine Learning Approach to House Price Estimation. *New Trends in Mathematical Sciences* 6(4), 165-171.

Wang, X., Li, K. & Wu, J. (2020). House Price Index Based on Online Listing Information: The Case of China. *Journal of Housing Economics* 50, 1-12.

World Bank (2015). *Stocktaking of the Housing Sector in Sub-Saharan Africa. Challenges and Opportunities*. Washington, D.C.: The World Bank.

World Bank (2020). *Purchasing Power Parities and the Size of World Economies. Results from the 2017 International Comparison Program*. Washington, D.C.: The World Bank.

World Bank (2021). *World Development Indicators*. Washington D.C.: The World Bank.

Yan, Z. & Zong, L. (2020). *Spatial Prediction of House Prices in Beijing Using Machine Learning Algorithm*. Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence, pp. 64-71.

Yoshino, N. & Helble, M. (2016). *The Housing Challenge in Emerging Asia*: *Options and Solutions.* Tokyo: Asian Development Bank Institute.

Yilmazer, S. & Kocaman, S. (2020). A Mass Appraisal Assessment Study Using Machine Learning Based on Multiple Regression and Random Forest. *Land Use Policy* 99 104889.

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable Selection for Hedonic Model Using Machine Learning Approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, 107(3), 293-306.

Zhang, B., Li, W., Lownes, N., and Zhang, C. (2021). Estimating the Impacts of Proximity to Public Transportation on Residential Property Values: An Empirical Analysis for Hartford and Stamford Areas, Connecticut. *International Journal of Geo Information* 10(44), 1-11.

Zhong, H. and Li, W. (2016). Rail Transit Investment and Property Values: An Old Tale Retold. *Transportation Policy* 51, 33-48.

Annex 1: Representativeness of the scraped data on the housing market

Country	Population (2019)	Average HH size	Number of Households	Number of scraped observations	Individuals using the internet (% of the population)	Percentage of all households covered in scraping	Slum Population, percent of urban population	Ratio of mortgages to GDP, percent
Albania	2,854,191	3.66	779,688	3,389	72.24 %	0.43	n.a	12.2
Costa Rica	5,047,561	3.20	1,579,421	10,376	80.53 %	0.66	4	15.93
Morocco	36,471,769	4.77	7,646,128	10,737	84.12 %	0.14	9	23
South Africa	216,565,318	6.28	34,500,045	107,652	68.20 %	0.31	26	16.15
Pakistan	58,558,270	3.86	15,155,778	91,904	17.00 %	0.61	40	0.23

Note: Overview on representativeness of the scraped house prices of the formal market. Numbers on average household size are drawn from countries' latest household survey. Data for total population are drawn from World Bank (2022).

Annex 2: Robustness: Results of Rigorous De-Duplication

	Random Forest-based Prediction						OLS-based Prediction				
Country	Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE		Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE
Albania	64,747.53	1.05	0.99	0.30	0.21		57,899.49	1.05	0.99	0.29	0.27
Costa Rica	111,008.21	1.00	1.02	0.32	0.26		117,282.53	1.03	1.12	0.38	0.31
Morocco	57,186.95	0.99	1.05	0.31	0.33		47,841.95	1.01	0.88	0.43	0.42
Pakistan	22,990.29	0.99	1.04	0.35	0.29		25,225.80	1.02	1.17	0.44	0.40
South Africa	63,728.53	1.01	1.05	0.30	0.30		85,469.41	1.12	1.28	0.54	0.55

Note: MAPE refers to mean absolute percentage error; local MAPE refers to the mean percentage error based on predictions accuracy of apartments sized between 50 and 60 square meters; local ratio refers to the ration between observed and predicted values for apartments sized between 50 and 60 square meters.

Annex 3: Alternate Outlier Removal

Albania

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Number of observations		3,389	3,376	3,064
Price (USD)	Median	97,619.05	97,619.05	98,809.52
Price (USD)	Mean	137,498.50	137,483.8	126,646.6
Square meter	Median	97.00	97.00	98.00
Square meter	Mean	116.73	115.84	109.37
Price per square meter	Median	1,046.57	1,047.62	1,066.68
Price per square meter	Mean	1,227.50	1,219.54	1,152.60
Price per square meter	IQR	508.32	505.95	493.19
Price per square meter	SD	2,886.30	2,804.12	526.90
Number of bedrooms	Mean	1.98	1.98	1.99

Costa Rica

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Number of observations		10,376	10,202	9,879
Price (USD)	Median	200,000.00	200,000.00	200,000.00
Price (USD)	Mean	293,751.60	295,211.70	267,586.4
Square meter	Median	181.00	181.00	180.00
Square meter	Mean	347.11	234.17	220.94
Price per square meter	Median	1,156.72	1,162.28	1,156.25
Price per square meter	Mean	2,744.14	1,393.17	1,314.42
Price per square meter	IQR	673.14	665.61	639.94
Price per square meter	SD	80,189.03	3,869.17	730.45
Number of bedrooms	Mean	.	.	.

Morocco

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Number of observations		10,737	10,734	10,131
Price (USD)	Median	104,395.60	105,494.50	104,395.60
Price (USD)	Mean	271,361.30	234,233.80	168,450
Square meter	Median	95.00	95.00	94.00
Square meter	Mean	188,59	133.68	117.90

Morocco (cont'd)

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Price per square meter	Median	1,119.25	1,117.21	1,130.10
Price per square meter	Mean	3,071.55	1,677.39	2,250.78
Price per square meter	IQR	932.10	903.46	925.03
Price per square meter	SD	33,458.45	7,570.71	14,499.67
Number of bedrooms	Mean	2.65	2.64	2.61

Pakistan

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Number of observations		107,652	107,524	103,847
Price (USD)	Median	83,934.34	83,934.34	83,934.34
Price (USD)	Mean	156,756.10	154,420.40	140,740.9
Square meter	Median	151.76	151.76	151.76
Square meter	Mean	955.86	220.93	210.80
Price per square meter	Median	535.38	535.38	535.38
Price per square meter	Mean	634.36	633.16	627.51
Price per square meter	IQR	344.14	344.14	339.65
Price per square meter	SD	652.22	480.86	395.14
Number of bedrooms	Mean	3.86	3.86	3.87

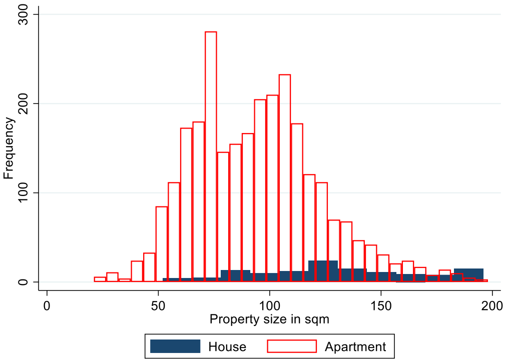
South Africa

Descriptive Statistic		*Original*	*Truncated*	*Winsorized*
		Data Set	Data Set	Data Set
Number of observations		91,904	88,374	70,608
Price (USD)	Median	95,588.23	93,286.45	105,498.7
Price (USD)	Mean	155,699.4	147,055	154,418
Square meter	Median	350	313	350
Square meter	Mean	2,081.1	555.01	803.73
Price per square meter	Median	360.38	387.17	357.06
Price per square meter	Mean	887.42	629.44	584.53
Price per square meter	IQR	614.62	624.15	599.32
Price per square meter	SD	9,302.60	853.02	700.54
Number of bedrooms	Mean	3.04	2.99	3.01

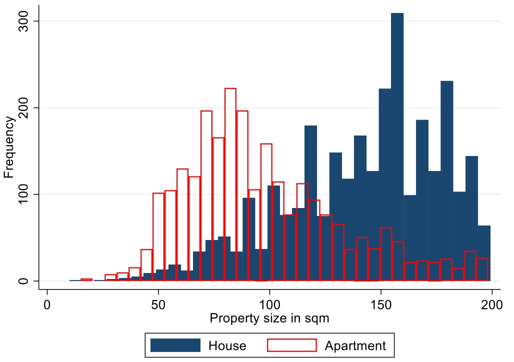
Note: The *original* data set contains the original set of all listings. The *truncated* data set retains listings that contain sale prices, size data, and whether the property is an apartment or house and truncates the data based on sqm<9 or sqm >3,000 and Price < $US 5,000 or Price > $US 5,000,000. The *winsorized* data set removes the first and 99th percentile of price and size; SD= Standard Deviation; IQR=Interquartile range

Annex 4: Frequency Distribution of Smaller Apartments and Houses

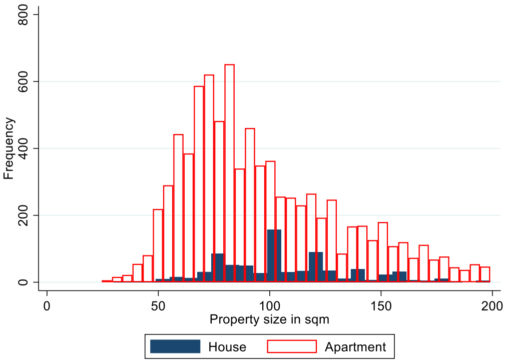
Albania



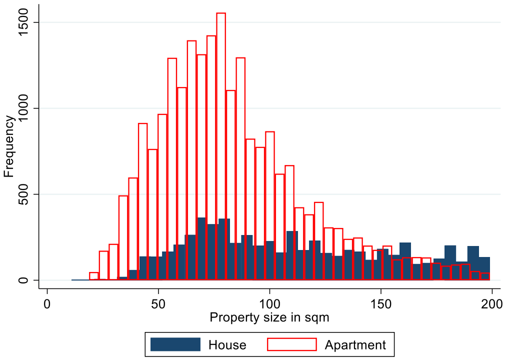
Costa Rica



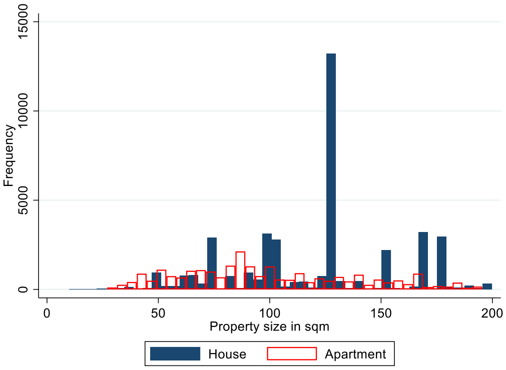
Morocco



South Africa



Pakistan



Annex 5: Robustness: Applying Different Splits

A) 75:25 Split

	Random Forest-based Prediction					OLS-based Prediction				
Country	Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE	Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE
Albania	64,151.96	1.00	1.05	0.29	0.24	64,128.05	1.00	1.03	0.29	0.34
Costa Rica	112,948.01	0.99	1.00	0.25	0.24	117,076.60	1.02	0.99	0.29	0.27
Morocco	50,540.15	1.00	1.04	0.24	0.27	47,497.29	1.02	0.99	0.41	0.43
Pakistan	21,953.53	0.99	1.13	0.31	0.27	24,590.69	1.01	1.26	0.44	0.51
South Africa	63,120.99	1.01	1.06	0.30	0.30	85,368.35	1.15	1.29	0.56	0.58

B) 90:10 Split

	Random Forest-based Prediction					OLS-based Prediction				
Country	Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE	Basic House Price *Apartment (Current USD)*	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE
Albania	64,225.80	1.01	1.14	0.26	0.25	56,673.70	1.00	1.10	0.28	0.26
Costa Rica	118,500.52	1.00	1.09	0.23	0.32	117,152.50	1.01	1.11	0.28	0.33
Morocco	51,580.91	1.00	1.04	0.24	0.26	47,689.99	1.03	1.21	0.45	0.52
Pakistan	22,309.16	1.01	1.12	0.30	0.27	24,613.56	1.03	1.27	0.44	0.53
South Africa	62,322.08	1.02	1.04	0.29	0.28	85,833.20	1.14	1.26	0.56	0.57

Note: MAPE refers to mean absolute percentage error; local MAPE refers to the mean percentage error based on predictions accuracy of apartments sized between 50 and 60 square meters; local ratio refers to the ratio between observed and predicted values for apartments sized between 50 and 60 square meters.

Annex 6: Robustness: Training of the RF Model in a Global Model

	Random Forest-based Prediction					
Country	Basic House Price *Apartment (Current USD)* Global Model	Basic House Price *Apartment (PPP$)* Global Model	Ratio predicted and observed price *(median)*	Local Ratio predicted and observed price *(median)*	MAPE	Local MAPE
Albania	55,923	148,600	1.07	1.10	0.28	0.24
Costa Rica	112,739	193,916	1.03	1.01	0.25	0.20
Morocco	56,405	137,148	1.07	1.10	0.29	0.27
Pakistan	24,330	101,633	1.09	1.29	0.36	0.45
South Africa	60,665	130,856	1.06	1.15	0.30	0.30

Note: MAPE refers to mean absolute percentage error; local MAPE refers to the mean percentage error based on predictions accuracy of apartments sized between 50 and 60 square meters; local ratio refers to the ratio between observed and predicted values for apartments sized between 50 and 60 square meters.