Team 012: BubbleWarn

Derek Cheng, Mahaveer Jain, Micah Jeng, Peter Hernandez, Victor Cerabone

# 1. Introduction

Asset-price bubbles are periods of time when the value of an asset increases dramatically and is followed by a comparable dramatic decrease in value. Housing market bubbles in particular have a large impact on the economy and tend to have longer lasting effects than other types of bubbles. Since the 2008 housing market crash, there have been economic debates over the potential of a similar bubble in the United States, however no clear data-driven warning system has been presented on the subject matter to reinforce these claims. According to Hans Lind (2008), an indicator system would be a set of characteristics such that if they are at hand during a period of quickly rising prices, then it increases the probability that prices will fall dramatically soon [1]. Dean Baker (2006) found that measuring the change in housing prices through the House Price Index (HPI) gave strong evidence of a bubble through the mid-1990s [2].Ole Martin Eidjord et al (2018) define HPI through a recursive regression analysis of select indicators [3]. The idea behind this project is to leverage common indicators during previous bubble periods and develop a bubble-warning system for the housing market assuming data on the identified indicators is available.

Along with house market bubble indications, it is also important to examine housing affordability to better understand the housing market. The index reflects not only the imbalance of supply and demand but can be tied to poor infrastructure planning, interest rate, and tax policy. Alex Lee (2020) found that the fear of missing out and the potential further rise in price can cause middle class families to take excessive risk and overextend their financial capacity [4]. Aaron Hedlund (2019) found that the rent-to-income ratio has grown from 25% to over 28% in the last decade[5]. The disproportionality between the increase in income and property prices coupled with any change in policy, supply chain and economic environment can lead to higher risk of default on a large scale which can be the source of the next bubble.

# 2. Problem Definition

A bubble can never be explained by a single factor but rather is the result of a complex culmination of factors [1]. In this paper, we are proposing a bubble warning system that aims to alleviate these debates through a data-driven analysis of the indicators and presents a speculative definition of a bubble-period. The housing affordability index will further provide insight into how it varies for a specific region.

## 2.1 Bubble Definition

Any definition of a bubble has edge cases, and one must take into account the relation between factors as well as the granularity of the analysis. This paper defines an asset-price bubble in such a way that it focuses only on the specific development of prices as denoted by the HPI and not on why the price developed in this way. Formally, a period of time will be regarded as a bubble if the HPI grows at least 50% during a 3 years period or at least doubles within a 5 year period [1].

## 2.2 Housing Affordability

Typically, housing affordability is derived as a ratio of housing price with income or ratio of mortgage payment with income as defined by Michael E. Stone [6]. This definition does not consider other factors such as down payment and therefore does not reflect the true cost of owning the house. We will examine various approaches to define affordability using available data and compare across different regions.

# 3. Proposed Method

One of the metrics used to define a bubble is the S&P CoreLogic Case-Shiller National Home Price Index that measures the change in sale price of single family homes across the United States [7]. The metric serves as a barometer of the U.S. housing market and the broader economy by tracking the purchase price and resale price of homes. The arrival of the mobile era and the widespread use of search engines enables a new way of estimating the HPI with the inclusion of the Google search volume index, (GSVI). The GSVI for a specified term is able to capture/measure the general public interest for a given topic. Wu and Brynjolfsson (2015) found that search queries can reveal the current housing trend and are well suited for predicting the future unit sales of housing [8]. Since Google search data became available in 2004 our analysis window will begin then and all other outstanding data sources will be filtered to exclude data prior to 2004. Our HPI regression analysis will be a complex culmination of the following factors: unemployment rate (UR), population (PO), interest rate (IR), number of housing permits authorized (HPA), and the Google Search Value Index (GSVI) for the search terms.

### 

## 3.1 Bubble Identification and Validation

Serving as a comparable metric, we juxtapose our bubble identification system output to the 2006/2007 housing market crisis of the United States. Moreover, in accordance with Ole M. Eidfjord (2018)’s analysis, we identify a set of benchmark states as an in-sample bubble identification test. Namely, Arizona, California, Florida, Maryland, Massachusetts, Nevada, New Hampshire, New Jersey, and Rhode Island serve as strong bubble states due to their large appreciation of HPI in 2006 and depreciation in 2007. Meanwhile the following states are considered weak bubble states: Alabama, Arkansas, Indiana, Iowa, Kansas, Nebraska and Oklahoma according to Lynch, Sharon (2008) due to the lack of strong depreciation of HPI at the end of the bubble in 2007 [9]. This test will serve as an effectiveness test of our bubble identification system.

## 3.2 Regression Analysis of HPI

To define the bubble using time-series data, we will refresh the housing price index using the linear regression and Error Correction Model (ECM). We begin with a linear regression of the HPI denoted below as Eq (1).

(1)

At time *t* and state *s*, *HPIs,t* represents the Housing Price index, *ꞵi* represents the regression coefficient for the corresponding variable, *HPIs,t-1* is the HPI at the previous time step t-1. *URs,t* is the unemployment rate, *POs,t* is the population. *IRt* is the average 30 year mortgage interest rate, *HPAs,t* is the number of housing permits authorized, and *GSVI(X)s,t* is the Google search volume index for the search term “X” which is keyword related to housing market.

In case if we find the terms (time series variable) are cointegrated, we will use ECM to further interpret the relationship.

The following steps will be taken:

1. Selection of independent variables using partial F-test and correction of seasonality and inflation using the ARIMA model. Perform variable transformation.
2. Drawdown analysis, which identifies the highest fall from the housing price peak, will be used to identify the potential bubble for a region. The duration of sharp changes will be defined as the bubble period and the rest will be treated as a non-bubble region.

**3.3 Housing affordability:**

Quan Gan (2009) examined and refined the affordability definition by breaking it down into 3 different levels [10]. Purchase affordability considers a household's capacity to borrow enough funds for a purchase. Repayment affordability considers the burden imposed on a household of repaying the mortgage. Income affordability simply measures the ratio of house prices to income. To combine all parameters to reflect true affordability, we have decided to use the ratio between Affordability Limit and ratio of house price and gross household income. A ratio higher than 1 reflects houses deemed unaffordable. The Affordability Limit (AL) captures the first two information (repayment and purchase affordability) and it is defined as:

where *α* represent the fraction of income towards mortgage payment, *ꞵ* the fraction of house price towards down payment, *N*the loan terms (in year) and*i* the interest rate. We defined HTI as the ratio between house price and gross household income, and it should always be lower than the Affordability Limit (HTI < AL).

## 3.4 Information Visualization

The project team chose to create the visualizations using Tableau desktop software and will publish the results on the site <https://public.tableau.com>. Our decision was based on research by Nair L. et al. (2016) , which found that creating interactive visualizations is more effective at providing insights to the user compared to traditional static visualization methods [11].

As far as granularity of the analysis is concerned, the housing bubble data is analyzed at the state level along with housing affordability at metropolitan areas across the United States and visualized using Tableau software for clarity.

# 4. Experiments

## With our proposed method, our project team has set up various experiments testing feasibility of our approaches. These include different time series ranges, limited by the columns chosen, and exploration of different models to best capture the time series data.

### 4.1 Data Collection/Integration

We plan to dynamically collect data from various web sources such as the Google Trends API and Zillow API. For our experiment, we chose to collect a static dataset with a confined date range for analysis and comparison of models from sources such as the U.S. Census Bureau, Federal Housing and Finance Agency, and the Federal Reserve Bank of St. Louis [12,13,14]. Each dataset had to meet certain criteria for it to be usable in our experiment. With the exception of interest and unemployment rates, data must be collected at the state level within the United States, uniquely identifiable via either a CBSA or FIPS code that denotes the appropriate state, and gathered at a yearly interval.

For the analysis of affordability, we explored various data sources and examined the compatibility before finalizing the data source. For housing prices, Zillow research provides the historical data of housing price index (ZHVI) across many metropolitan areas and mortgage rates for a given period [15]. American Community Survey (ACS) provides extensive data for earnings at occupation level **[**16]. Finally, we utilized Crosswalk data to link Zillow regions and federally defined regions for counties and metro areas.

**4.2 Data Analysis**

Bubble identification system: In the case of datasets not collected at the state level, we merge the data by year. In all other cases, we merge the data by either the CBSA or FIPS code to create one ordered time series dataset for analysis.UNDER CONSTRUCTION.

Affordability limit: ACS data (summary files) are challenging to use due to 3.5 million data attributes stored across various files in sequence. The meta information about attributes, sequence, and geographical information are stored in separate files which require additional effort to understand the ACS data standards. Python and the requestmodule are mainly used to access the data from various sources.

At the end, we merged the Zillow data and ACS data to compute the ratio between Affordability limit (AL) and HTI for metropolitan areas across the US.

We decided to use Python for data gathering, clearing, model selection, and performance evaluation.

**4.3 Data Visualization**

The main visualization on the final tableau dashboard will be a choropleth map outlining CBSA geographic regions. According to the US Census Bureau [12], they created CBSA codes to define metropolitan or micropolitan areas. Shown in Figure 1 is an experimental dashboard with each region on the choropleth map shaded according to the income ratio metric, however the housing price index will replace this metric in the final versions. This experiment also confirmed Tableau can automatically generate CBSA boundaries [17].

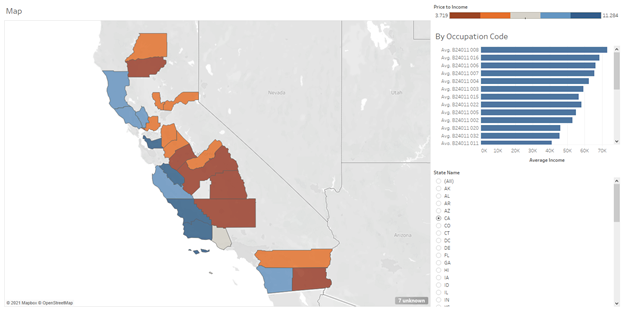


Figure 1. Choropleth Map on Tableau

# 5. Conclusion

The most significant project development has been to evaluate whether any indices can be used to detect the housing bubble. This is trending positive, however more evaluation is needed. We are focusing on implementing a linear regression model. The project scope is more focused compared to our proposal, and notable is that this warning system provides no insight into the duration of a bubble-period, only statistical confirmation that the specified region has entered a bubble. Further predictive analysis is recommended for such results. We currently project time and resources are sufficient. *UNDER CONSTRUCTION*

## Team member activities and contribution:

|  |  |  |
| --- | --- | --- |
| **Group Member** | **Previous** | **Current** |
| Derek Cheng | Data cleaning and analytical modeling, report organization | Analytical modeling, report organization, and report writing |
| Mahaveer Jain | Data acquisition and preprocessing, analytical modeling and literature survey | Data acquisition and preprocessing, analytical modeling, and report writing |
| Micah Jeng | Data processing, visualization and report | Data source collection, data cleaning, and report writing |
| Peter Hernandez | Data collection and model development, visualization | Data source collection, data cleaning, and report writing |
| Victor Cerabone | Tableau, presentation, and video | Tableau, presentation, video, and report writing |
| Tsz Kin Chan | Tableau and data preprocessing | Left project team |

After the loss of a team member, responsibilities are re-distributed and everyone has contributed equally.

# 

# References

1. Hans Lind. Price Bubbles on the Housing Market: Concept, theory and indicators. In International Journal of Housing Markets and Analysis 2(1):78-90
2. Dean Baker, Is the Housing Bubble Collapsing? 10 Economic Indicators to Watch. 2006. In RePEc, January 2006.
3. Ole Martin Eidjord, Are Oust. Can Google Search Data be Used as a Housing Bubble Indicator?
4. Alex Kae Lun Lee, Jothee Sinnakkannu, Sockalingam R. Ramasamy(2020). Housing bubble, affordability and credit risk to banks: a Malaysian perspective. Journal of the Asia Pacific Economy; v.26.
5. Aaron Hedlund (2019). [Housing Affordability: Trends, Consequences, and Policies](https://aaronhedlund.github.io/policy/CGO_housingaffordability.pdf).
6. Michael E. Stone. What is housing affordability? The case for the residual income approach. 2010.
7. S&P CoreLogic Case-Shiller Home Price Indices Methodology, <https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-corelogic-cs-home-price-indices.pdf>
8. Wu, L. and Brynjolfsson, E. (2015) The Future of Prediction How Google Searches Foreshadow Housing Prices and Sales, Chapter in NBER book Economic Analysis of the Digital Economy, editors Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, editors (p. 89 - 118).
9. Lynch, Sharon (October 2, 2008).. ["Metro U.S. Home Prices Fall on Higher Foreclosures"](https://www.bloomberg.com/apps/news?pid=20601213&sid=aAaW9UAEs8Dk&refer=home). [Bloomberg](https://en.wikipedia.org/wiki/Bloomberg_L.P.). Retrieved October 10, 2008.
10. Quan Gan and Robert J. Hill. Measuring Housing Affordability: Looking Beyond the Median. In Journal of Housing Economics, Volume 18, Issue 2, June 2009, Pages 115-125. 2009.
11. Nair L., Shetty S., Shetty S. (2016), Interactive visual analytics on Big Data: Tableau vs D3.js, Journal of e-Learning and Knowledge Society, v.12, n.4, 139-150. ISSN: 1826-6223, e-ISSN:1971-8829
12. Bureau, US Census. “Core-Based Statistical Areas.” Census.gov, 8 Oct. 2021, <https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html>.
13. Federal Housing Finance Agency. “House Price Index Data”, <https://www.fhfa.gov/>
14. Economic Research, Federal Reserve Bank of St. Louis. “Unemployment Rate”, <https://fred.stlouisfed.org/>
15. Zillow, Housing Data,<https://www.zillow.com/research/data/>
16. [American Community Survey](http://www.census.gov/programs-surveys/acs/data/summary-file.html) (ACS) Survey summary file.
17. Location data that tableau supports for building map views. Tableau. (n.d.). Retrieved November 4, 2021, from <https://help.tableau.com/current/pro/desktop/en-gb/maps_data.htm>.