Harnessing Zillow Data to Understand Coastal Property Valuation

Jack Bienvenue

Department of Statistics

University of Connecticut

February 13, 2025

Introduction Residential properties in coastal areas are often highly coveted and the degree of settlement in coast-adjoining counties in the United States is unproportionate to their small contribution to the land size of the country. (Landry et al., 2022) Some researchers, like Gourevitch et al. (2023), argue real estate prices in coastal communities are critically overpriced as markets do not acknowledge discounting for burgeoning risk factors. Climate-intensified natural disasters and sea-level rise threaten coastal communities, yet coastal real estate markets retain significantly higher average property values, and those average valuations continue to rise. (McNamara et al., 2024) In this project, we will conceive of, build out, and implement a novel data science workflow which attempts to replicate this finding in McNamara et al. (2024) using open data, namely from the popular real estate listing website Zillow.

Specific Aims This analysis will center around using random sampling to evaluate whether average valuations for coastal property are growing disproportionately in comparison to non-coastal properties. The expectation will be to observe this phenomenon occurring as it has been observed in literature. (McNamara et al., 2024) This result is of interest as it appears to be in contrast to the expected market response of pricing intensifying anthropogenic climate change risks in coastal property valuations. (Fuerst and Warren-Myers, 2021)

Data Description For this analysis, data will be obtained solely from open sources to ensure that reproducibility is accessible to interested parties.

The data from Zillow can be accessed for free through NASDAQ Data Link's open data resources. The data is free to use, so long as a free account is made to access an API key. The Zillow data is an extremely large set of retail real estate transactions. Zillow additionally has an internal system of defining geographic sales areas, so this will need to be taken into consideration for the encoding of coastal markets, which are not an explicit entry in any table provided by Zillow. The data for the main table of transactions lists:

1. the unnamed primary key representing the transaction ID,
2. *indicator_id* representing the type of geographic region whether it be a county, zip code, or Zillow-defined neighborhood,
3. *region_id* representing a five-digit code for Zillow's internal geographic encoding,
4. *date* representing a transaction date, and
5. *value* representing the home sale price.

Two auxilliary tables define the indicator types and the regions. In addition to the Zillow data, shapefiles from data.gov will be used to determine whether regions are adjacent to the coast.

Research Design and Methods Given the incredibly vast nature of the Zillow data, it will be necessary to employ randomized sampling in order to remain within time and computational constraints.

Random sampling will begin by generating 30000 random numbers within the range of number of home IDs in the Zillow dataset without replacement. While a large sample, this represents a small proportion of the total entries in the transaction data. Since data may be ordinal, before sampling, the next step will be to reset applicable indices and randomize the order entries in the large file before subsetting it.

After the random sample is drawn, mechanisms will be created to sort the region into a designation of coastal and non-coastal. This may involve a variety of sorting methods. For instance, landlocked states can be automatically applied to the non-coastal category. Extraction of city names, zip codes, or county names from the renaming set can be used for matching to a shapefile of US municipalities, zip codes, or counties to create the coastal or non-coastal designation.

Having constructed the desired dataset, time series visualizations of coastal vs non-coastal valuations can be created. From here, after curve-fitting, tests can begin to be made to examine differences in the rates of valuation change over intervals of time. Selection of models for these regressions and tests will be made following the consultation of Dr. Gu and Dr. Schifano.

Discussion Climate risks are of concern to coastal communities and residents for a number of reasons. These includes risks induced by low probability, high impact severe weather events and the constant threat of sea level rise. It is foreseeable that a market acknowledgment of climate-related risks could be devastating to home valuations in coastal areas. In addition, insurance carriers may be reluctant to continue to carry coastal homeowners, and coastal customers may experience rate hikes to their insurance. (Peterson et al., 2024)

Since we generally expect markets to be efficient and to price in risks, including climate risks, for asset valuations, the finding in McNamara et al. (2024) that coastal property values are still on the rise at a rate surpassing non-coastal properties is surprising. This analysis can serve as an independent confirmation or challenge of this result.

This analysis is potentially limited by the lack of control involved in the construction of the sample. Zillow's anonymization of data makes it hard to determine what the relationship is between a property and nearby coast. The dataset we construct will be agnostic to this relationship. For example, waterfront properties may have high prices, but a property in the inland part of a coastal region may not be subject to market dynamic of coastal property. In addition, inland properties may experience different reactions to climate risk. For example, low-lying areas of inland Florida could react differently to climate risks as compared to inland New England. While acknowledging this issue, random sampling may allow for some of the adverse effects of this lack of granularity to be averted.

Core elements of Data Science

Programming - All progress on this assignment will require robust scripting. Python in a Quarto environment will be utilized in order to create script outputs which are well organized and annotated.

Data management - This project will require the harnessing of data of varying types and from various sources. Organization will be of the essence, and Git and Github will be used throughout the course of the project to ensure version control and ease of access.

Data analysis - Analysis will be central to the process of gathering results from the data science workflow we construct. We will implement hypothesis testing techniques to identify if there have been intervals of statistically significant unproportionate gains in coastal real estate valuations when compared to non-coastal properties.

Data visualization - Visualization will be a key part of this analysis, and elements of time series visualization will be central to understand trends in valuations. Spatial visualizations will be created when appropriate to display the random sample of properties. Descriptive visualizations will be used to understand distributions of data attributes. Visualizations will be shared with the class during presentations.

Data ethics (collection and use) - In terms of the collection process, this data is solely gathered from open data sources from trusted entities. Real estate data is anonymized to the level of the region ID, so that particulars of individual properties are not sensitive data. Data used in this study are true, representative of reality, and reported in an unbiased manner to the best of my knowledge. Data use practices in this analysis are not inclined to endanger the rights or well-being of any individual or organization.

Conclusion Developing novel processes for answering research questions is a core aspect of what is at the heart of being a data scientist and researchers. This project proposes an novel, open-data statistical data science approach to independently evaluate conclusions made in published academic work on the topic of coastal real estate valuations. This topic is of great importance as coastal communities are subjected to heightening climate risks in the wake of anthropogenic climate change.

References

Fuerst, F. and Warren-Myers, G. (2021), "Pricing climate risk: Are flooding and sea level rise risk capitalised in Australian residential property?" *Climate Risk Management*, 34, 100361.

Gourevitch, J. D., Kousky, C., Liao, Y., Nolte, C., Pollack, A. B., Porter, J. R., and Weill, J. A. (2023), "Unpriced climate risk and the potential consequences of overvaluation in US housing markets," *Nature Climate Change*, 13, 250–257.

Landry, C. E., Turner, D., and Allen, T. (2022), "Hedonic property prices and coastal beach width," *Applied Economic Perspectives and Policy*, 44, 1373–1392.

McNamara, D., Smith, M., Williams, Z., et al. (2024), "Policy and market forces delay real estate price declines on the US coast," *Nature Communications*, 15, 2209.

Peterson, J., Hill, A., Dandridge, J., Kousky, C., and Jones, D. (2024), “The Coastal Property Insurance Crisis,” *Env’t L. Rep.*, 54, 10443.