Spring 5-10-2024

Assessing the Economic Consequences of Sewage Overflows on California Housing Values

Nathalie M. Nyanga
nmnyanga@usfca.edu

UNIVERSITY OF SAN FRANCISCO

Flushed Away: Assessing the Economic Consequences of Sewage Overflows on California Housing Values

by

Nathalie Nyanga*

Supervised by Dr. Robizon Khublashvili

Submission : May 10, 2024

*Email: nmnyanga@dons.usfca.edu

Abstract

Sewage overflows (SSOs) represent a critical environmental challenge in California, and the impact of SSOs on housing prices is a vital issue. The study aims to determine how the frequency of sewage overflow events affects housing price fluctuations and conduct an event study. Employing linear, lasso, ridge, and random forest regression analyses, the study predicts the severity of overflow. Notably, lasso regression emerges as the most effective model for predicting housing prices in the event of an SSO occurrence. Additionally, the event study analysis provides insights into the factors influencing housing prices.

Table of Contents

1 Introduction

Sewage overflows (SSOs) represent a critical environmental challenge in California, endangering public health and potentially causing overlooked effects on the state's housing market, particularly amid housing shortages. The overflows are the unintended release of untreated sewage into the environment due to obstructions, infrastructure failures, or overwhelmed sewer systems, attracting attention from policymakers and residents alike. Nevertheless, their impact on housing markets is under-explored.

Our investigation initially presumed a negative impact of sewage overflow events on housing prices, aligning with common assumptions regarding environmental hazards. However, upon deeper consideration and pre-study analysis, we formulated the null hypothesis, that sewage overflow events have no significant impact on housing prices. This shift in hypothesis underscores the complexity of environmental influences on housing markets and emphasizes the need for analysis.

The study seeks to bridge this gap by examining the correlation between the frequency and severity of sewage overflows across significant California counties—namely Alameda, San Mateo, San Joaquin, Marin, and Sacramento—and the fluctuations in housing prices over time. By considering property characteristics, we aim to unravel the intricate relationship between environmental factors and housing market dynamics, providing invaluable insights for homeowners and policymakers alike.

The research employs a comprehensive approach that includes spatial econometric analysis, mapping overflow occurrences by volume, and using linear, ridge, Lasso, and random forest models to predict the effects of overflows on housing prices and conduct an event study. The study examines literature, methodological approaches, empirical findings, and conclusive insights to develop effective strategies for mitigating the impacts of sewage overflows on housing market resilience and public well-being.

2 Literature Review

In exploring the interplay between sewage overflows and housing prices, the analysis of existing literature was vital; it helped frame the scope and approach for the paper. The analysis involved synthesizing different studies to analyze further the effects of environmental factors on the housing market and the surrounding environment.

Chiaburu and Biessmann (2024) investigated using interpretable time series models for predicting critical water levels in combined sewer overflows (CSOs). They investigated using interpretable time series models for predicting critical water levels in combined sewer overflows (CSOs). The study discussed how the increasing frequency of extreme weather has strained the sewer system in Ahr, Germany. They described how CSOs become overwhelmed due to heavy rain, leading to releases that pollute nearby water bodies. The paper proposes that interpretable time series models can predict critical water levels in CSOs, allowing for better management and reduced environmental impact. The researchers used these models to predict water levels during heavy rain events. They aimed to identify water levels that could lead to overflows and develop strategies to redirect excess flow within the sewage network before overflows occur. The findings suggest that modern interpretable time series models have the potential to be a viable solution for predicting wastewater level rise caused by heavy rainfall. It could significantly improve wastewater management and prevent environmental pollution from CSOs. However, the models require robust software and hardware for real-world application in sewer systems.Additionally, the models have limitations in tracking overflow levels of sensors within the system crash. They are also predictive and cannot account for human interventions. Further research is necessary to address these limitations and improve reliability.

The impact of wind farms on prices in Poland is a vital study that analyzes the effects of environmental factors on the housing market. According to Torzewski M. (2016), it analyzes existing literature on wind farms' effects on property prices within Poland. The study analyzes the current and older studies within Poland, locally and internationally. Poland experienced increased wind energy production from 2009-2010, becoming the country's most significant renewable energy. Despite wind energy growth, there are fewer studies on the causal effects of wind farm development on housing prices in Poland. The study also found that international studies found adverse effects on nearby housing if created; however, the results were less consistent. The study found that some international findings are not directly applicable to Poland due to location: wind farm placement strategies might differ between Poland and other countries—public perception: Public attitudes towards wind farms could vary across regions. The findings indicated the need for increased research in Poland to assess the impact of wind farms on property values, considering relevant factors in the Polish housing market. The research emphasizes the lack of conclusive evidence on wind farms' effects on housing markets concisely across the market. While international studies suggest negative impacts in Poland, it still needs to be clarified, requiring increased analysis of the knowledge gap.

According to Harrison, D., Rubinfeld, D. L., it delves into a novel approach to estimate how much people value clean air by analyzing data from the housing market. The core idea is that people would be willing to pay more for houses situated in areas with better air quality. The study's findings support this notion, revealing that people pay a premium for houses in cleaner-air locations. This preference for clean air is even more pronounced among higher-income households, suggesting that they value this environmental amenity more. It is important to note that the strength of this effect depends on how the equation used to model housing prices is structured. The authors acknowledge that the sensitivity of the results to the model specification highlights the need for careful consideration when designing such studies. The research offers a unique perspective on how housing market data shed light on people's preferences for environmental quality, specifically clean air. The findings suggest that people are willing to pay more to live in areas with cleaner air, and this willingness to pay is extreme for wealthier households. However, the study underscores the importance of robust methodologies when analyzing these relationships. The strength of the results depends on the specific model used to analyze housing prices. It raises concerns about the generalizability of the findings if the model structure changes. Housing price data may not capture all factors influencing people's choices. Other amenities or location features could influence the price premium besides air quality. The study establishes a correlation but does not definitively prove causation. People with higher incomes may choose cleaner areas for other reasons, not just air quality, that influence housing prices. Overall, the research by Harrison and Rubinfeld offers a valuable contribution by demonstrating the potential of using housing market data to understand environmental preferences. However, the limitations highlight the need for further research with more robust methodologies to solidify the cause-and-effect relationship between air quality and housing prices.

Through these analyses, we have deepened our understanding of the factors influencing property values and refined our approach to include a broader range of variables. This comprehensive review not only reinforced the relevance of our study but also provided a robust framework for exploring the economic impacts of environmental and municipal services on real estate markets.

3 Methodology

The study aims to determine how the frequency and occurence of sewage overflow events effect with fluctuations in housing prices within the five California counties by using the literature review as a guide to address gaps within prior studies. To approach the research the hypothesis was established which were the null hypothesis established sewage overflows occurrence has no impact on housing values within counties while the alternative hypothesis establishes sewage overflows negatively impact housing values within the surrounding area.

The research employed mixed methods, including spatial analysis, linear, lasso, ridge, and random forest regression, as well as event study spatial analysis, to comprehensively assess the impacts of sewage overflows on housing prices. Data for the study were collected from the California Open Data Portal, the US Census Bureau, County Open data for California Counties (specifically, Marin, Alameda, Sacramento, San Joaquin, and San Mateo), and Zillow. The dataset comprises 1,132 observations spanning from 2009 to 2023, containing information on home prices, physical housing characteristics, and sewage overflow attributes (such as spill volumes in net gallons of human waste). The five California counties chosen were selected for their prevalence of severe sewage overflows within the state. These counties possess different geographic and economic contexts, contributing to the diversity of the representative sample. The characteristics examined include: the price of the home where the sewage overflow occurred, the age of the home, the number of bedrooms, the number of bathrooms, interior sqft (the amount per square meter where the actual home building occupies), lot sqft (the amount of square footage occupied), the event variable (representing sewage overflow occurrence if it exceeds a threshold of 1000 net gallons of sewage), and spill volume (the volume of sewage associated with the property).

The analysis employed linear, lasso, ridge and random forest regression was used to predict how fluctuations in prices were affected by sewage overflows. The main model is:

$$\text{Price} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Bedroom} + \beta_3\text{Bathroom} \quad + \beta_4\text{Lot sqft} + \beta_5\text{Interior sqft} + \beta_6\text{Spill vol}$$

.

For the event study, the years 2016 to 2020 were selected due to the passage of legislation in 2016 in the state of California called California Code, Water Code - WAT § 13377. This legislation, an extension of the National Clean Air Act, provides the framework for enforcement actions against entities that violate discharge permits, leading to stricter enforcement against overflow occurrences. Fines or quicker corrective actions during occurrences and facility shutdowns are mandated for private and public entities violating the laws, necessitating compliance with federal regulations aimed at preventing pollution and maintaining water quality. The event study analyzed the effects before and after the legislation to determine if sewage overflows had an impact on housing prices fluctuations. The equation used for this analysis was:
$Price = \beta_0 + \beta_1\text{Marin} + \beta_2\text{Sacramento} + \beta_3\text{San Joaquin} + \beta_4\text{San Mateo} + \beta_5\text{Age} + \beta_6\text{Bedroom} + \beta_7\text{Bathroom} + \beta_8\text{Lot Sqft} + \beta_9\text{Interior Sqft} + \beta_{10}\text{Spill Volume}$

The spill volume which represents sewage overflows is threshold is a 1000 gallons which is starting threshold for medium spill volume which are caused by larger blockages, pump failures or over flows moderate rain they have an impact compared to smaller spills. Spatial Analysis was used to map sewage overflows distance and occurrence at exact location in the county. The Python geopandas library was employed to map the overflow sites, highlighting the different regions geographically and the severity of the overflows within the regions where they occurred. Overall, Python was utilized to carry out each method for each section or model created.

4 Results

The study undertakes a comprehensive examination of the relationship between sewage overflow events (SSO) and housing price fluctuations in Alameda, San Mateo, San Joaquin, Marin, and Sacramento counties of California. We will analyze summary statistics, examine visualizations, compare models, and conduct an event study analysis to investigate this relationship further.

4.1 Summary Statistics

We will first examine summary statistics derived from the datasets to provide foundation of central tendencies within the data. The table below includes the important summary statistics of variables analyzed.

Table 1: Average Summary Statistics of Variables by County

County	Age	Bed	Bath	Lot_sqft	Interior	Spill	Event	Price
Alameda	67.43	2.89	2.16	6350.69	1803.15	14554.41	109	1400902.82
Marin	42.19	3.48	2.64	13943.41	2541.14	1022.75	104	1733118.25
Sacramento	22.05	3.51	2.54	20151.35	4999.64	191.66	811	783703.55
San Joaquin	54.35	3.16	2.44	6582.27	2151.81	1044.02	102	555560.57
San Mateo	69.67	3.29	2.81	14172.26	2714.26	2370.31	180	2349852.63

Table 1: Notably, within had the highest sewage overflow events occurrence. The highest average home prices and age were within San Mateo. The largest average spill volume was within Alameda county.

4.2 Visual Analysis

We use the correlation matrix and the distribution of sewage overflows

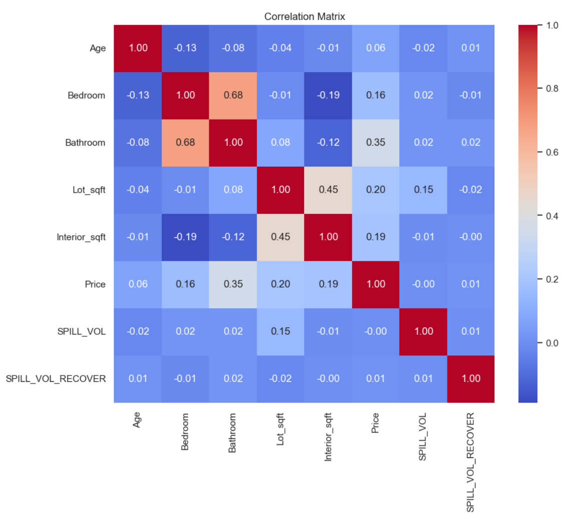

Figure 1: Correlation Matrix

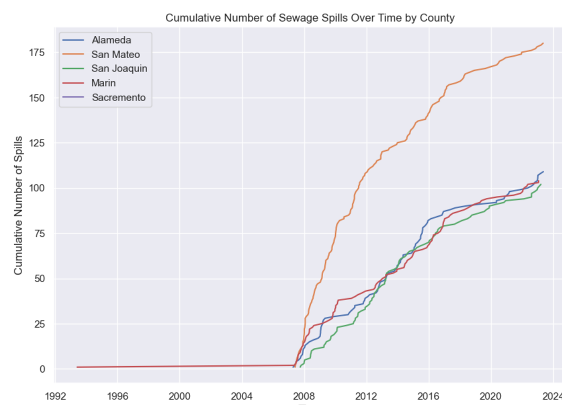

Figure 2: Cumulative Sewage Spill's

Figure 1 displays a correlation matrix. It shows distribution of correlation among variables red being positive (closer to 1) and blue being lowest correlation (closer to -1). An example is bedrooms and Bathroom shared the highest positive correlation at 0.68 while Interior sq-ft and Bathroom have the negative correlation being the lowest at -0.12. The figure 2 shows the cumulative distribution of sewage overflows. It was found that Alameda had the highest cumulative volume of sewage spill volumes per net gallon. the other counties show increase in spills however the trend is much steeper indicating a slower rate of increase. The data is reflecting as an increase in spills is either increasing or improvement within reporting and recording spills overtime.

4.3 Spatial Analysis

The maps below show the distribution of sewage overflows; they serve as visual representations of the overflows observed in each county analyzed. To facilitate our analysis, we established a mapping system that categorizes the severity of overflows based on volume. In this system, orange indicates the most severe cases, corresponding to 100,000 gallons of new sewage, while yellow represents mid-level overflows at 10,000 gallons of net sewage. Green denotes 1,000 gallons of net sewage, and blue indicates the least severe instances, with less than 100 gallons of sewage overflow. Through this mapping system, we were able to discern key patterns and trends in sewage distribution across counties. In the subsequent discussion, I will briefly outline the significant findings from each county's analysis

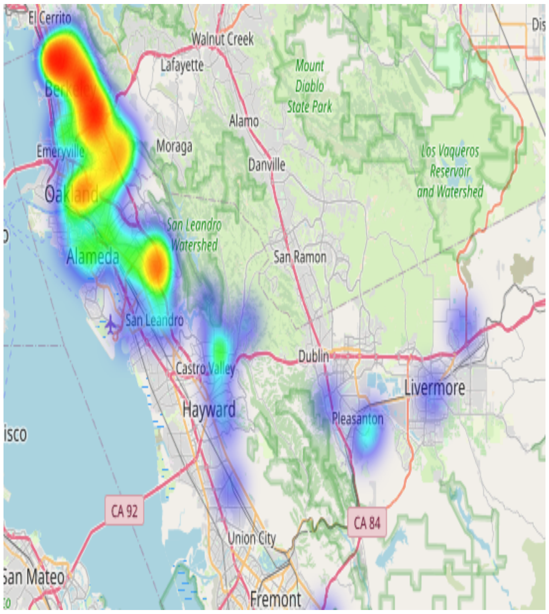

Figure 3: Alameda

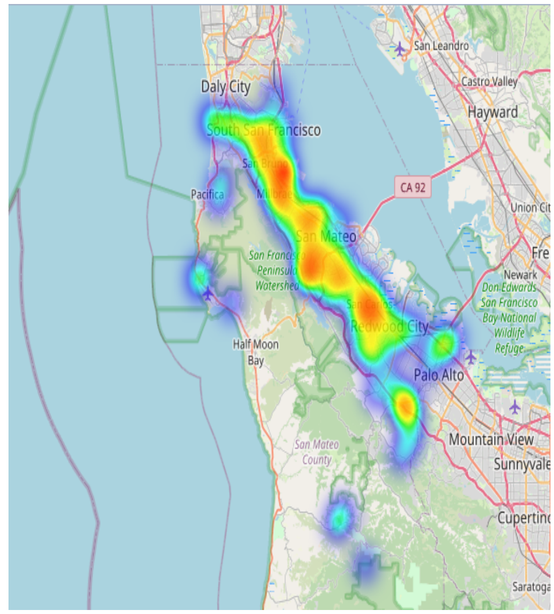

Figure 4: San Mateo

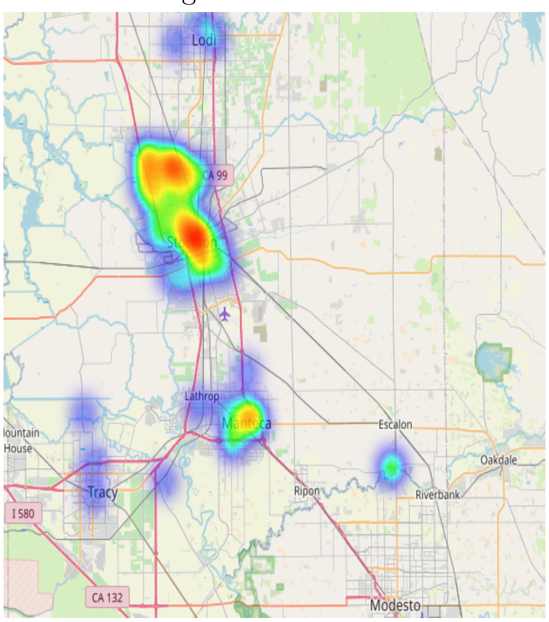

Figure 5: San Joaquin

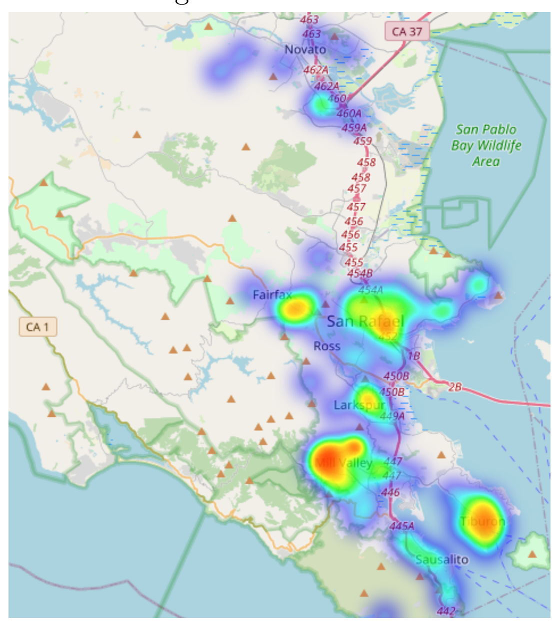

Figure 6: Marin

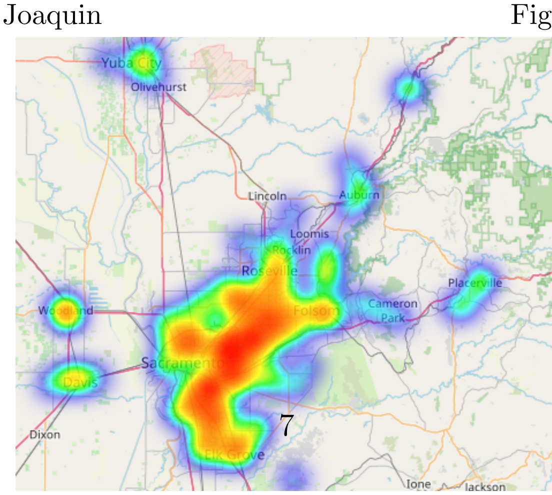

Figure 7: Sacramento

The significant findings from analysis were: In Marin County, the highest concentration of sewage overflows clustered around Mill Valley, with fewer occurrences noted in Novato and Sausalito. Notably, the prevalence of overflows in Mill Valley, despite its status as a high-net-worth area, underscored the reality that affluent regions are not immune to sewage infrastructure challenges. Moving to Sacramento County, urban areas like Sacramento and Folsom exhibited the highest concentration of overflows, contrasting with lower occurrences in Auburn. This pattern highlights the critical role of densely populated areas in contributing to sewage system strain. San Joaquin County displayed a similar trend, with Manteca and Stockton experiencing the highest concentration of overflows, while Tracy reported fewer incidents. This disparity reflects the challenges faced by regions with older infrastructure and lower housing values, particularly in urban areas like Stockton. San Mateo County's analysis unveiled San Bruno as a hotspot for overflows, attributed in part to its proximity to water bodies and transportation hubs like San Francisco Airport. Lastly, Berkeley emerged as a focal point for sewage overflows in Alameda County, emphasizing the vulnerability of hilly regions near water bodies to sewage system strain. Overall, the findings show that sewage overflows transcend socioeconomic boundaries, with geographic and infrastructural factors playing big roles in their occurrence. These insights can inform targeted infrastructure investments and mitigation strategies to address sewage overflow challenges comprehensively across the regions.

4.4 Machine Learning Analysis Predicting Housing Prices

To approach predicting changes in housing prices, various regression models were employed, including linear, lasso, ridge, and random forest regressions. The analysis will begin with interpreting the results of the linear regression, followed by an overview of all models.

Table 2: OLS Regression Results

	Coefficient	Std. Error	Z Score	P> $\|z\|$
Intercept	13.5419	0.022	605.597	0.000
Age	0.2535	0.024	10.394	0.000
Bedrooms	0.0891	0.039	2.291	0.022
Bathrooms	0.3477	0.034	10.352	0.000
Lot Sqft	0.0110	0.055	0.200	0.841
Interior Sqft	0.1621	0.092	1.761	0.078
Spill Volume	0.0068	0.009	0.721	0.471

The analysis revealed that bedrooms, bathrooms, and age were statistically significant at the 5 % level, meaning they have a statistically relevant impact on housing prices. Conversely, spill volume, representing sewage overflows in gallons, was insignificant, suggesting it has no impact on prices in this model. For each unit increase in a home's age, we expect a housing price increase of 0.2535 (coefficient value). Lot square footage and interior square footage, however, were not statistically significant (p-value > 0.05), indicating they may not have a significant influence on housing prices within this model.

Table 3: Linear, Lasso, Ridge and Random Forest Regression Results.

Table 3: Regression Performance

Model	$\mathbf{R^2}$	**MSE**
Linear	0.668	0.446
Lasso	0.393	0.113
Ridge	0.434	0.324
Random Forest	0.625	0.526

Among the models evaluated, the Lasso regression model emerges as the most effective in terms of predictive power. It has a R^2 value of 0.393, the Lasso model explains approximately 39.3% of the variability in housing prices, showcasing its strong explanatory ability. Additionally, the model exhibits a relatively low mean squared error (MSE) of 0.113, indicating accurate predictions of housing prices on average. In contrast, while the Linear regression model achieves a higher R^2 value of 0.668, it exhibits a higher MSE of 0.446, suggesting larger prediction errors compared to the Lasso model. Despite the higher R^2 value of the Linear model, the emphasis on predictive accuracy favors the Lasso regression model. Therefore, Lasso regression is the best at predicting housing prices if a sewage overflow event occurred.

4.5 Event Study

The event study in this research aims to investigate the influence of legislative measures, notably California Water Code - WAT § 13377, on sewage overflow events and their effects on housing prices in selected California counties (Alameda, San Mateo, San Joaquin, Marin, and Sacramento). By incorporating variations in property characteristics and analyzing how this relationship varies across counties. This comprehensive approach aims to provide insights into housing market dynamics and legislative influences, ultimately revealing the interconnected of sewage overflows and housing prices in California.

Table 4: Event Study

Variable	Coefficient	Std. Error	t-statistic	P-value
Intercept	13.1430	0.351	37.412	0.000
COUNTIES				
Marin	-0.0384	0.309	-0.124	0.901
Sacramento	-0.9584	0.254	-3.768	0.000
San Joaquin	-0.9769	0.293	-3.332	0.001
San Mateo	0.5486	0.284	1.929	0.056
Event	-0.0435	0.221	-0.197	0.844
Age	-0.0021	0.003	-0.677	0.500
Bedroom	0.1004	0.066	1.530	0.129
Bathroom	0.2923	0.076	3.867	0.000
Lot_sqft	6.076e-06	2.02e-06	3.007	0.003
Interior_sqft	3.31e-07	4.04e-06	0.082	0.935

The event study analysis provides insights into the factors influencing housing prices in the studied counties. The intercept suggests that the base price of a property, assuming zero bedrooms, bathrooms, lot square footage, etc., is approximately 13.1430. Coefficients for each county (Marin, Sacramento, San Joaquin, San Mateo) reveal differences in housing prices compared to a reference county, with San Mateo exhibiting higher prices. However, sewage overflow events, represented by the Event variable, do not appear to have a statistically significant impact on housing prices, as indicated by the non-significant coefficient (-0.0435, p ¿ 0.05). Instead, variables such as Age, Bedroom, Bathroom, Lot sqft, and Interior sqft show significant associations with housing prices, with coefficients indicating the estimated change in housing prices for a one-unit increase in each respective variable. Overall, while property characteristics and county location play significant roles in determining housing prices, the influence of sewage overflow events may not be statistically significant in the studied areas.

5 Conclusion

The findings from both the event study and machine learning analysis provide valuable insights into the relationship between sewage overflow events and housing price fluctuations across multiple California counties. While our initial assumption before the study was that sewage overflow events would negatively impact housing prices, our hypothesis stated no discernible impact. However, upon analysis, the null hypothesis, indicating that sewage overflow events, as indicated by spill volume, do not demonstrate a statistically significant effect on housing prices in the Alameda, San Mateo, San Joaquin, Marin, and Sacramento Counties, is rejected. This unexpected outcome challenges our preconceived notion and underscores the importance of empirical investigation in revealing the true nature of the relationship between environmental factors and housing market dynamics.

Interestingly, despite our initial expectation of a negative impact, the data suggests that sewage overflow events do not exert a discernible influence on housing prices. Instead, variables such as the number of bedrooms, bathrooms, lot size, and interior square footage emerge as more influential determinants of housing prices. Moreover, the application of various machine learning models, particularly the lasso model, yielded robust predictive accuracy, further emphasizing the need for data-driven insights in understanding the complexities of the housing market.

However, several limitations inherent in the study should be acknowledged. Firstly, the reliance on historical data for sewage overflow events may not fully capture the dynamic nature of these occurrences, potentially leading to underestimation or overestimation of their true impact. Moreover, while the machine learning models provide valuable insights, they are subject to inherent assumptions and limitations, including the need for careful hyperparameter tuning and the risk of overfitting.

To strengthen causal inference, several approaches could be considered. Instrumental Variables (IV) could be used if there are external instruments that influence sewage overflow but not housing prices directly. Propensity Score Matching (PSM) could match properties based on their likelihood of experiencing a sewage overflow to similar properties that didn't, allowing for comparison of their price changes. Difference-in-Differences (DiD) could be employed if pre- and post-event data are available, helping to isolate the impact of sewage overflows by comparing the change in housing prices over time between affected and unaffected areas.

Despite these limitations, the study contributes to our understanding of the complex interplay between environmental factors and housing market dynamics. Future research could explore alternative methodologies for assessing the impact of sewage overflow events, incorporate additional variables to capture nuanced variations in housing prices, and further investigate the potential policy implications of these findings. By addressing these limitations and building upon the insights gained, future studies can continue to advance our understanding of the broader implications of environmental events on real estate markets.

6 Appendix

6.1 References

1. Ashfield Chad. (2023, September 11). Concern over extra sewage from new housing after King's Mill Reservoir pollution; Sewage could be examined amid fears new Ashfield housing may affect water quality. *Mansfield, England.*

2. California Legislature. (n.d.). Assembly Bill No. 1588. Retrieved from https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200AB1588

3. Chiaburu, T., & Biessmann, F. (2024). Interpretable Time Series Models for Wastewater Modeling in Combined Sewer Overflows. *Journal Name, Volume(Issue)*, Page numbers.

4. Environmental Protection Agency (EPA). (n.d.). Sanitary Sewer Overflows (SSOs). Retrieved from https://www.epa.gov/npdes/sanitary-sewer-overflows-ssos

5. FindLaw. (n.d.). California Water Code Section 13377. Retrieved from https://codes.findlaw.com/ca/water-code/wat-sect-13377/

6. Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*(1), 81-102. https://doi.org/10.1016/0095-0696(78)90006-2

7. Meilinger, V., & Monstadt, J. (2022). FROM THE SANITARY CITY TO THE CIRCULAR CITY? Technopolitics of Wastewater Restructuring in Los Angeles, California. *International Journal of Urban & Regional Research, 46*(2), 182–201. https://doi.org/10.1111/1468-2427.13014

8. Petrie, B. (2021). A review of combined sewer overflows as a source of wastewater-derived emerging contaminants in the environment and their management. *Environmental Science and Pollution Research, 28*(25), 32095–32110. https://doi.org/10.1007/s11356-021-14103-1

9. Torzewski, M. (2016). The Impact of Wind Farms on The Prices of Nearby Houses in Poland: a Review and Synthesis. *Real Estate Management and Valuation, 24*(2), 13-24.

10. Wambecq, T., Kroll, S., Van Assel, J., & Delgado, R. (2023). CSO Generator—A Parsimonious Wastewater Quality Model for Combined Sewer Overflows. *Water, 15*(19), 3424. https://doi.org/10.3390/w15193424

6.2 Appendix A: Visualizations

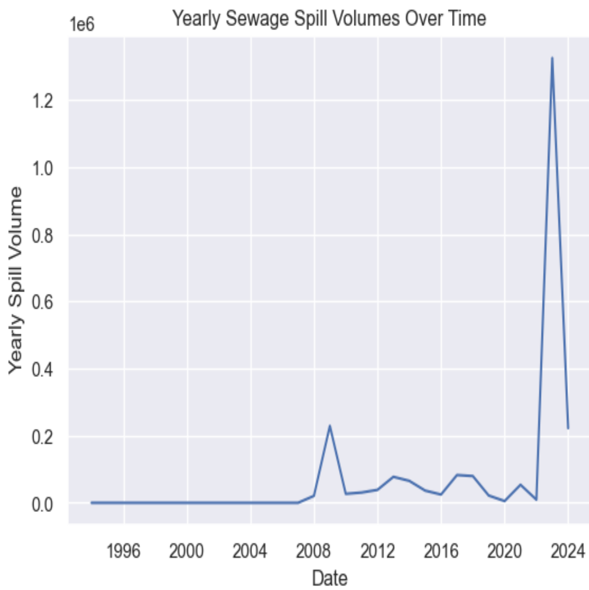

Figure 9: Yearly Sewage Overflow

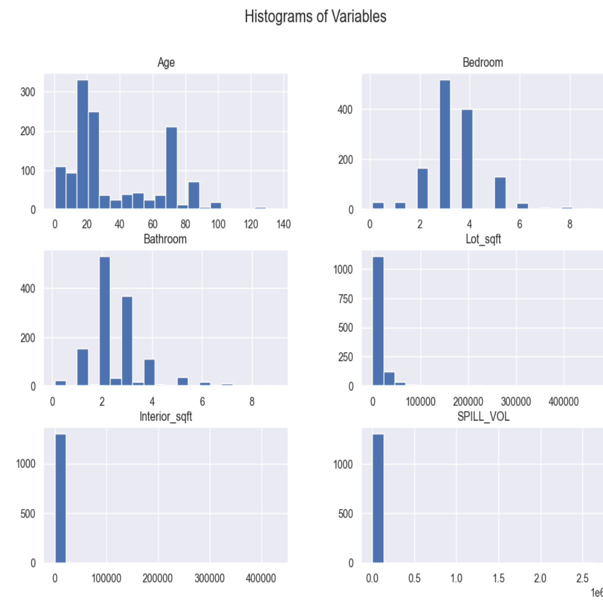

Figure 10: Summary Statistics

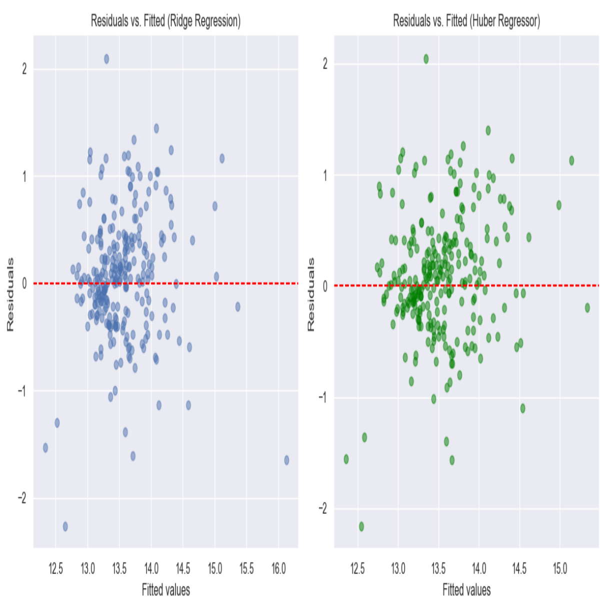

Figure 11: Ridge Regression Residuals

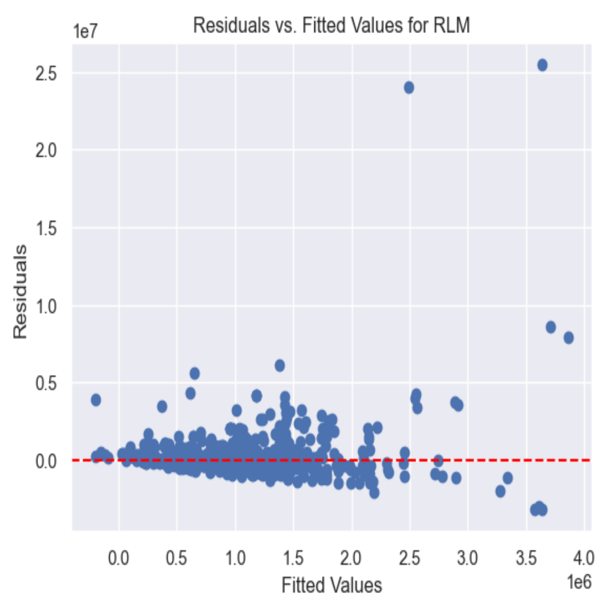

Figure 12: Linear Regression Residuals