S&DS 625 Capstone Project - Spatio-Temporal Modeling of California Real Estate Investment Activity

Matthew Ross

12/16/2024

Abstract

Opportunistic real estate investors look to find markets in which the current value of properties differ from their intrinsic value, and profit on purchased properties when the market corrects. While some characteristics of undervalued properties are more tangible and property specific, such as unusually low asset prices, positive market conditions such as the overall economic health of a region, or characteristics of the local population play a part in the market correction assumption of an investor. This project aims to uncover what regional market characteristics drives institutional investor deal volume in California over time, with the goal of being able to predict regions before they receive future deal flow. The study employs a variety of datasets, using real estate investment deals from Preqin (to approximate Deal Volume, the outcome variable), and socio-economic predictors from the U.S Census Bureau, Bureau of Labor Statistics, Zillow, and the Internal Revenue Services of the United States. A spatio-temporal model is used to predict Deal Volume (across all types of investors) in a 6-month time period, which accounts for the various economic predictors, but also any spillover effects from nearby towns, as well as the time effect of previous deal flow on future deal flow. We find that greater Unemployment Rate within an area and a greater Federal Funds Rate strongly signal lower overall deal flow, and areas with high Home Value and a greater population with Welfare Benefits and Private Health Insurance signal higher overall deal flow. The time effect of previous deal flow is a strong predictor of future deal flow, and spillover effects between towns are significant within a 20km radius. While predictive power of the model is weak, we show that San Diego could be considered to be a favorable investment in 2024. Overall, results from this model show that macroeconomic effects are primarily responsible for an areas investment, but there are interesting societal characteristics that seem to be desired (either purposefully or inadvertently) by investors.

Introduction

Real estate investment firms can be split into two categories, opportunistic and core.

A core real estate investor will look to invest in assets that are in prime locations, have stable cash flows, and are generally considered to be low risk. These investors are looking for a steady return, and are willing to pay a premium for the stability of the asset. As an investor in this firm, you will expect to have general exposure to the real estate market as whole.

An opportunistic investor, on the other hand, is looking for assets that can be purchased at a price that is less than their intrinsic value, and has the potential for high returns as the market eventually corrects itself or through improved operational management of the property. These market dislocations can be due to a variety of reasons: distressed sellers looking to dump assets, a neighborhood's poor reputation driving down prices, operational mismanagement, etc.

These investors are willing to take on more risk in order to achieve this highreturn, particularly through betting on the probability of market correction, and the time it sill take to do so. Skilled investors will be able to achieve significant outperformance of the general real estate market.

The playbook for an opportunistic investor often entails them parachuting into new, relatively niche markets where there appear to be market dislocations. But what are the signals for investors that indicate that an area is home not only to mispriced or mismanaged assets, but also is a market which will correct itself to drive returns over a long time period? Are these indicators purely macroeconomic, or does there exists socioeconomic data which can be used to find favorable markets? The identification of these predictors is the question this analysis seeks to answer.

It is difficult to distinguish solely based on deal metrics and location what is an "opportunistic" or "core" investment. This project aims to uncover what factors drive overall deal volume, and then attempts to interpret how these could apply to an opportunistic investor. Therefore, the direction of coefficients should not be considered a conclusive signal for solely an opportunistic investor.

As David Swensen outlines in his seminal work, *Pioneering Portfolio Management*, institutional investors like endowments and foundations achieve outsized returns by being able to parse which real estate managers, especially opportunistic ones, are able to outperform the market. With many real estate managers of late boasting extensive data science capabilities, the questions this report seeks to answer will hopefully show if there actually are meaningful investment signals that can be captured through data.

When thinking about this problem, one cannot ignore two primary factors: time and location. Time is of the essence when identifying undervalued markets - with billions of dollars at the command of investors looking for these opportunities. Therefore, the effect of previous time periods of investment has significant impact on future near term investment. Next, location cannot be considered to be independent. Characteristics of a city, such as its economic health, the health of its population, are often correlated with that of nearby areas, meaning that real estate investment in one city can spillover into another.

To combine the influences of space and time alongside the predictors in our model, we employ the INLA (Integrated Nested Laplace Approximation) spatio-temporal model. The INLA model allows us to estimate the posterior distribution of fixed effects (like our economic predictors), and random effects (spillover effects, time), and therefore understand their contribution of these effects to Deal Volume. Specific building of the model is discussed further in the paper. Given the computational complexity of the model, we limit our analysis to the state of California - which we believe to be comprehensive given its size and diversity of markets.

While the most tangible data points in real estate investment are often property specific, this information is often not publicly available, and costly to obtain from brokers or private sources. Therefore, we look to find macroeconomic and socioeconomic, publicly available predictors that real estate investors have historically found (either purposefully or inadvertently) to signal a good market for entry. We detail the data sources and predictors below. Based on data availability, we limit our analysis from January 2010 to June 2023.

1. Preqin Data

- **Description**: This dataset contains deal-level information for the state of California, including details such as deal name, date, type, location, and asset attributes. Data ranges from early 1990 to October 2024.

Primary Variables Used:

- **Asset Type:**
  - Type of assets include Residential, Retail, Industrial, Office, Niche, Land, Hotel, and Mixed Use
- **Asset City:**

  - City where the asset is located
- **Deal Volume (Outcome Variable):**
  - Aggregate number of deals per city in 6-month period

2. Bureau of Labor Statistics (BLS) Data

- **Description**: This dataset includes employment metrics (labor force, employment, unemployment) by geographic area. Data is collected monthly.

Primary Variables Used:

- **Unemployment Rate:**
  - Percentage of unemployed individuals in the labor force.

3. Federal Reserve Data

- **Description**: Time series data for federal interest rates. Data is collected monthly.

Primary Variables Used:

- **Federal Funds Rate:**
  - Interest rate at which depository institutions lend reserve balances to other depository institutions overnight.

4. Zillow Data

- **Description**: This dataset contains Zillow Home Value Index (ZHVI) for single-family homes and condos. According to Zillow, the ZHVI reflects the typical value for homes in the 35th to 65th percentile range. Data is collected monthly.

Primary Variables Used:

- **Home Value:**
  - Median home value for single-family homes and condos.

5. ACS (American Community Survey) Data

- **Description**: Contains demographic and economic metrics for cities across California, such as median household income, health insurance coverage, job status, and more. Data is collected annually.

Primary Variables Used:

- **Median Household Income:**
  - Median income of households in a city.
- **Welfare Benefits:**
  - Total population receiving public assistance (Foodstamp/SNAP benefits, Public Assistance)
- **Health Insurance:**
  - Total population with private health insurance coverage, and total population without any health coverage.
- **Government Benefits:**
  - Total population receiving government benefits (Social Security, SSI, Veterans, Medicare, etc.)
- **Job Type:**
  - Total population by job type. Grouped by "White Collar", "Blue Collar", "Other" within the analysis.
- **Transportation Method:**
  - Total population in possession of a car, and population of those taking other primary transportation methods (public transport, carpooling, etc.).
- **Mean Travel Time to Work:**
  - Average travel time to work for the population in a census area.
- **Work from Home:**
  - Total population working from home.
- **Unpaid Family Workers:**
  - Total population working as unpaid family workers. Defined as an individual who works without pay for 15 or more hours a week in a family owned and operated business.

6. Internal Revenue Services (IRS) Opportunity Zone Data

- **Description**: Spatial data for designated Opportunity Zones in California. The IRS designates an Opportunity Zone as "an economically-distressed community where new investments, under certain conditions, may be eligible for preferential tax treatment." Data is only available as of 2019.

Primary Variables Used:

- **Opportunity Zone:**
  - Binary variable indicating whether a city is designated as an Opportunity Zone as of 2019.

The structure of the report is as follows: Section 1 provides data exploration and visualization of the Preqin dataset and the economic predictors, showing how they affect Deal Volume. Section 2 details the modeling and analysis approach, where we identify macroeconomic key predictors using an INLA spatio-temporal model, and assess the model's predictive power. Finally, Section 4 concludes the report, providing recommendations and ideas for future work.

1. Data exploration and visualization

Visualizing the Preqin Dataset

We begin our analysis by examining the Preqin data, in order to understand how Deal Volume changes over time within California.

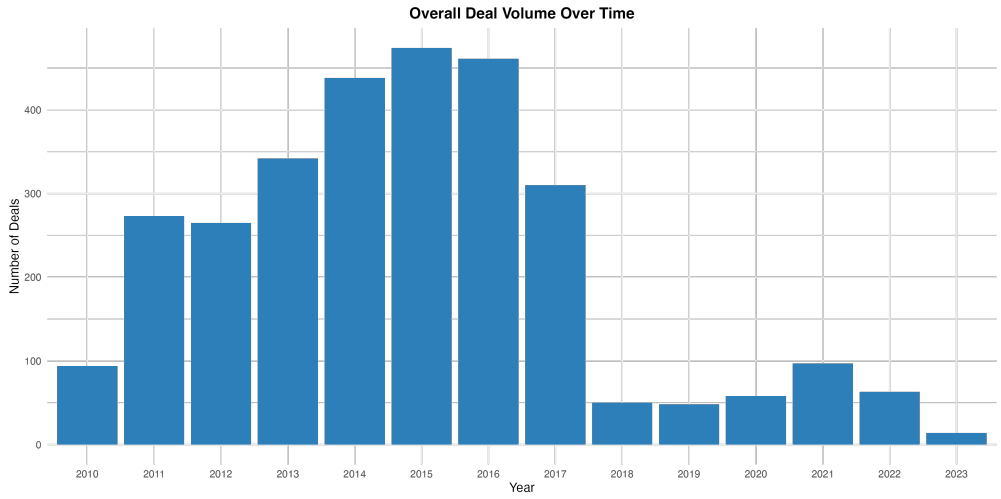

Figure 1: Overall Deal Volume over Time

While one would expect Deal Volume to decline sharply during the onset of the COVID-19 pandemic, it is evident from the figure that Deal Volume in California reduced sharply in 2018. While we could not find any glaring factor that would cause this reduction, we know that this behavior is likely not due to missing data, since for other states within the Preqin dataset, Deal Volume remains stable until 2020.

Therefore, some potential reasons for this could be: general uncertainty surrounding interest rates in 2018 curbing homeowner demand, bottoming out of housing affordability in 2018 (California Association of REALTORS, 2018), or uncertainty surrounding the voting of Proposition 10, which would have allowed local governments to enact rent control on any type of rental housing.

We can further visualize this geographically as follows:

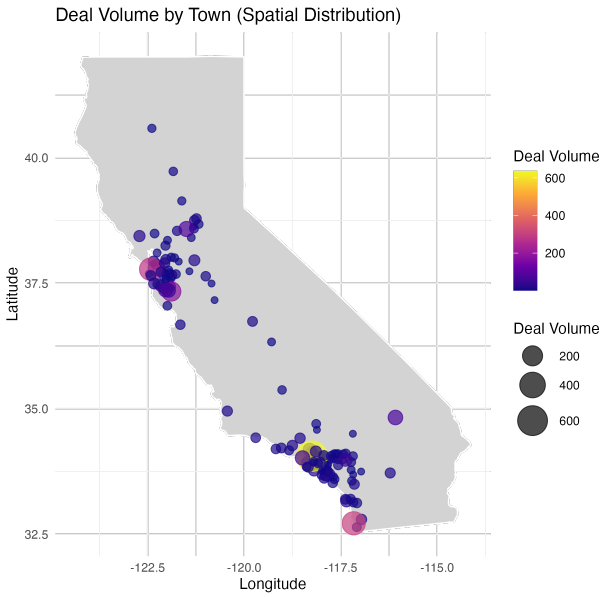

Figure 2: Total Deal Volume Map

Here, we can see that the majority of our data is split between the northern and southern ends of the state, and largely coastal. From this, we can see that the largest cities by Deal Volume are Los Angeles, San Diego, and San Francisco. The close proximity of townships to eachother further emphasizes our need to utilize spatial correlation in our model.

Moving on, we look to see the distribution of deals across asset class, and if there are types of real estate that are particularly popular within California.

Figure 3: Deal Volume by Asset Type

Here, the majority of deals are made up of residential and office spaces. This is not surprising, given that these are the most common types of real estate in California. However, it is possible that post COVID-19 office will be a less popular asset type, given the rise of remote work. We will examine how work-from-home population in an area affects Deal Volume in the modeling portion of this analysis.

We also look to see if this Deal Volume is focused in any specific cities. The heatmap below examines the 25 cities (there are 136 unique towns within the data) with the most deals in California for readability.

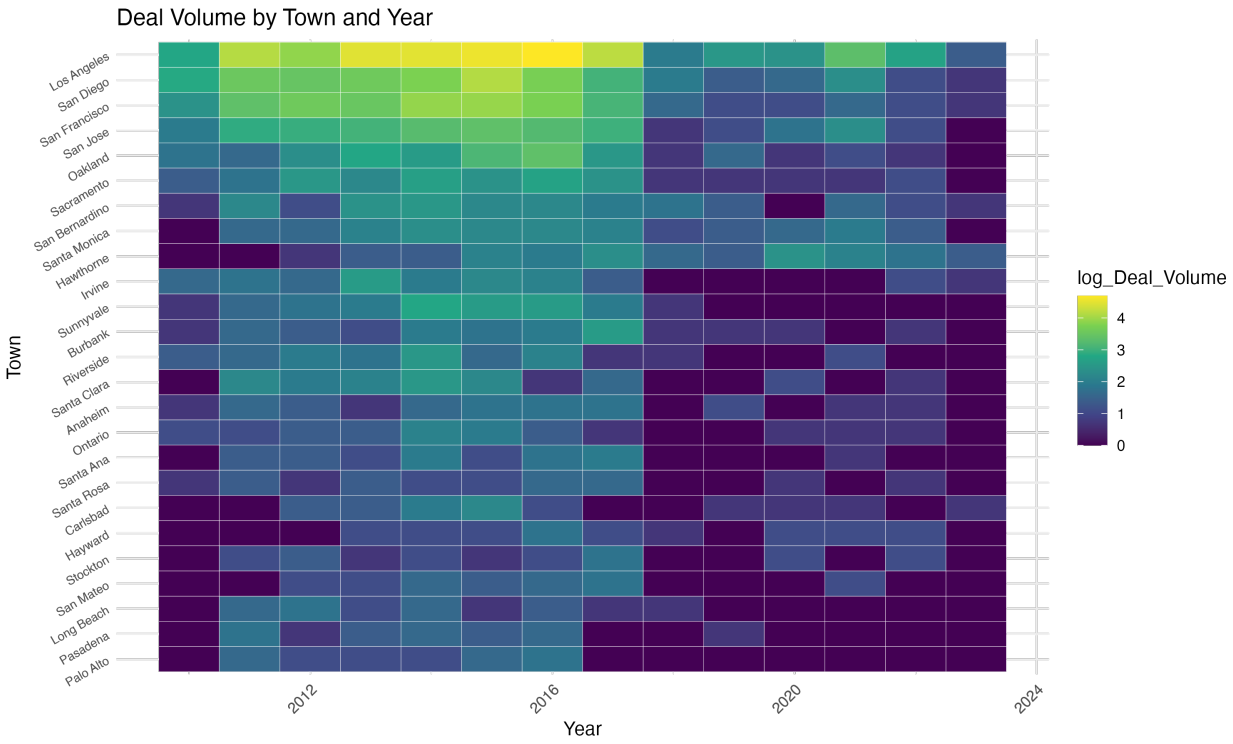

Figure 4: Deal Volume, by City and Year

Here, we can see that the vast majority of deals are located in Los Angeles, San Diego, and San Francisco, seen in the yellowish colors in the top. The visualization aligns with previous reasoning of the peak Deal Volume occurring pre-2018. In this graph, we do see that towns like Hawthorne have seen an increase in deal activity over the past 5 years.

We can look deeper at the deal flow within Los Angeles over time with the distributions of deals by address using the following gif file, here.

From the gif file, it is possible to see that the majority of historical deals within LA have occurred in the downtown area, with some expansion outside of this area beginning in 2020.

Visualizing the Economic Predictors

We now move on to visualize the economic predictors that we will use in our model. First, we will examine their effects graphically. While we only include a 2 of the strongest predictors here, the rest of the predictors are located in the Appendix.

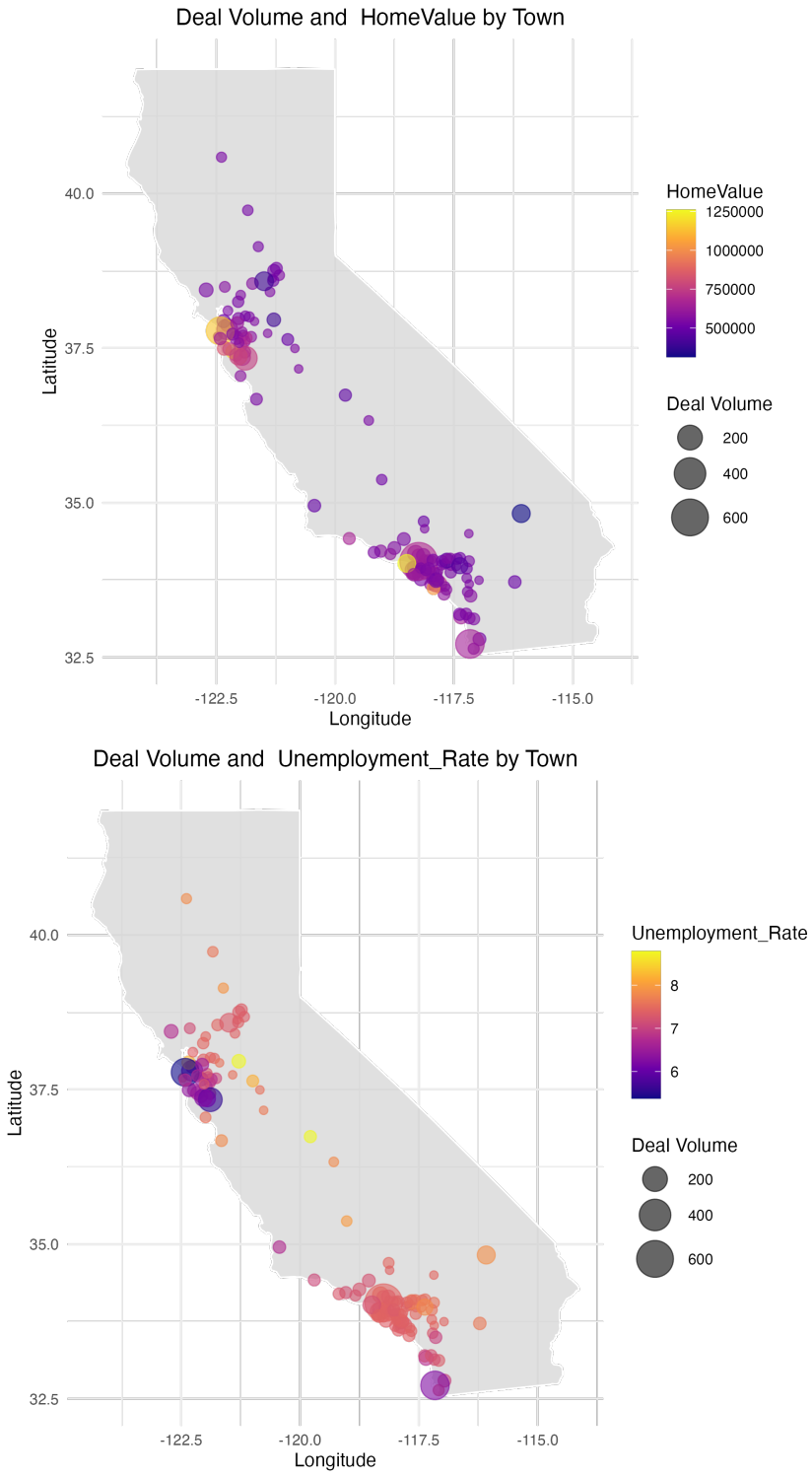

Figure 5: Scaeld Predictors Vs. Deal Volume

In these two plots we can see Home Value and Unemployment Rate by Town, and how they are potentially related to large Deal Volumes. For Home Value, we can see that the highest valued properties in California are in the Bay Area, and do command a sizeable Deal Volume. However, lower valued properties in the San Diego and Los Angeles area also command a sizeable Deal Volume. For Unemployment Rate, we can see

that the lowest unemployment rates are in the Bay Area, with higher rates in the South.

The relationship between Unemployment Rate and Home Value is clear (lower unemployment, higher home values). However, it is difficult to discern the relationship they have to Deal Volume.

For our last plot, we will observe scatterplots for each strong predictor against Deal Volume. We note that the predictors have been scaled, by subtracting the mean and dividing by the standard deviation of each predictor across all deals. In addition, for interpretability, the y axis is the log of the Deal Volume.

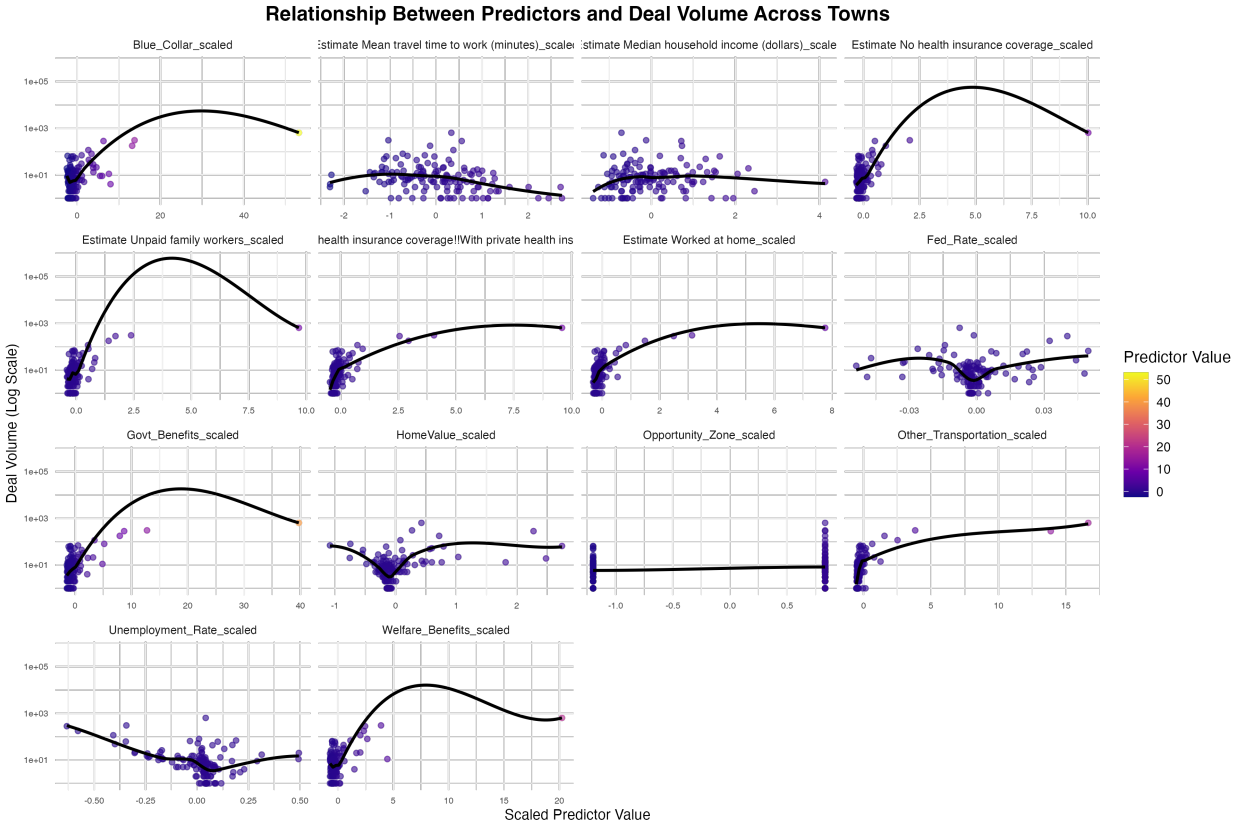

Figure 6: Unemployment Rate, Home Value by Town

From this plot, we can deduce several things. First, the existence of a single outlier often throws off what would be an approximately linear relationship to create an inverted parabola shape. This outlier is likely Los Angeles, and its position not along the linear trend suggests that for large areas, the relationship between predictors and log Deal Volume is not linear.

If we ignore the outlier, we can some relationships within the data. Notably, there appear to be approximately positive effects of Blue Collar population, no Health Insurance and Private Health Insurance population, Other Transport (i.e public transit usage), Work from Home population, Unpaid family worker population, and Government Benefits population on Deal Volume.

Unemployment Rate, Travel Time to Work appear to be negative contributors, with Federal Interest Rate, Home Value, and Median Household Income having a more difficult to discern relationship with Deal Volume.

Now that we have introduced our predictors, we will move on to the formation of our model.

2. Modeling/Analysis

We employ an INLA spatio-temporal model to predict Deal Volume for a town in a six month time period. Since we are modeling count data over a period of time, the model can be viewed as a Poisson regression model with a spatio-temporal component.

2.1 Modeling Details

Modeling Assumptions

1. **Poisson Distribution:** The model assumes that the number of deals in each town and period follows a Poisson distribution, meaning the variance equals the mean, and is solely count data.
2. **Linearity on the Log Scale:** Predictors enter the model linearly on the log scale (log-link function).
3. **Spatial and Temporal Structure:** The model includes spatial random effects, assuming towns closer together share similar deal patterns, and temporal random effects, assuming that the number of deals in one period is correlated with the next.
4. **No Excess Zeros or Overdispersion Unaccounted For:** The model assumes the Poisson distribution is appropriate, i.e., no significant zero-inflation or more variability than Poisson allows. Given the data contains sparse deal flow for many towns across long periods, we will explore the use of a zero-inflated Poisson model.

Observation, Predictors, and Outcome

- **Observations:** Each row represents a particular town and period combination.
- **Predictors (Columns of X):** These include scaled economic indicators (e.g., Fed_Rate_scaled, Unemployment_Rate_scaled, HomeValue_scaled), demographic measures (e.g., Estimate.Worked.at.home_scaled), and other socio-economic variables. Each predictor is a column in X, and is normalized by subtracting the overall variable's mean and dividing by its standard deviation.
- **Outcome (y):** The response y is the count of deals observed for that town and period.

The INLA model can be viewed as incorporating the spatial and temporal effects into a standard Poisson regresion model. Therefore, we can stylize the equation as:

$$\log(\lambda_{it}) = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \cdots + \beta_p X_{p,it} + u_{space,i} + v_{time,t}$$

where:

- λ_{it} is the expected Deal Volume for town i at time t.
- β_j are the fixed-effect coefficients for each predictor.
- $u_{space,i}$ is the spatial random effect for town i.
- $v_{time,t}$ is the temporal random effect for period t.

Interpretation of the Coefficients

A coefficient β_j corresponds to the expected change in the log of the deal count per unit increase in that predictor, holding others constant.

- **Positive β_j:** Increasing the predictor is associated with higher predicted Deal Volumes.
- **Negative β_j:** Increasing the predictor is associated with fewer deals.

Hyperparameters

The model includes hyperparameters associated with the spatial and temporal random effects:

- **Spatial Hyperparameters:**
  - **Range for spatial.field (θ_{range}):** This parameter indicates how far spatial correlation extends; a larger range means that towns farther apart still share some similarity in Deal Volumes. We choose to measure this in meters.
  - **Precision for Period (τ_{Period}):** Precision is the inverse of variance; higher precision indicates less variability in the temporal random effects.
  - **Rho for Period (ρ_{Period}):** An autoregressive parameter - when close to 1 this suggests strong temporal correlation, meaning that Deal Volumes in one period are highly related to those in the subsequent period.

Performance Measures

- **Fit Measures:**
  - **DIC (Deviance Information Criterion):** Provides a measure of model fit that accounts for model complexity. Lower values indicate a better fit.
  - **WAIC (Watanabe-Akaike Information Criterion):** Another fit measure that assesses predictive accuracy. Lower values are better.
- **Predictive Ability:**
  - **Residual Analysis:** Assessing residuals to identify patterns not captured by the model. We can also use this to test overdispersion.
- **Marginal Log-Likelihood:** Can be used for model comparison - higher (less negative) values generally indicate a better-fitting model.

Model Appropriateness

This Poisson model with spatial and temporal random effects is suitable if the data follow a count distribution with no severe overdispersion or zero-inflation. The presence of correlated measurements across towns and over time is not problematic given the inclusion of the spatial and temporal effects.

- **Appropriate If:**
  - The mean and variance of the deal counts are approximately equal.
  - Spatial and temporal correlations exist and are significant.

In our data, there are a high amount of zeroes. We therefore test a zero-inflated Poisson model to account for this. In addition, we will test a negative-binomial model to deal with potential overdispersion.

Ease of Interpretation

While the formulation of the INLA model is quite complicated and difficult to explain even to a technical audience, the interpretation of its results are fairly straightforward - they can be viewed like a Poisson GLM with fixed and random effects (which are spatial and temporal).

For a non-technical audience, the signs and magnitude of predictors are able to explain positive or negative effects on the outcome variable. While the spatial and temporal effects are more difficult to explain, they can be understood as the effect of nearby towns and previous time periods on Deal Volume - and can be explained fairly simply through the values of the hyperparameters θ_{range}, τ_{Period}, and ρ_{Period}.

2.2 Modeling Process

To reduce the initial predictor space and identify the most valuable predictors, Lasso regression was first employed to aid in variable selection. The remaining variables were then consolidated into different categores, like Government Benefits, Welfare Benefits, etc. to reduce multicollinearity.

To build the INLA model, the process detailed in (Wilke et. al, 2019, pg. 169) was employed. To capture spatial dependencies, we use an SPDE (Stochastic Partial Differential Equation) model. This model approximates a continuous spatial field (i.e, series of latitude and longitude coordinates from towns) by projecting it onto into a 2-dimensional triangular mesh. This mesh retains the geographic shape of California. A visualization of the mesh can be found in the Appendix. Then, we encode the temporal dynamics using a first-order autoregressive model.

The results from these models are simply fed in as random effects to the INLA model, alongside the fixed effects predictors discussed in the introduction.

Following this, an iteration over all predictor combinations within the INLA model was used, and the subset of predictors that achieved lowest WAIC was chosen as the best set.

2.3 Visualization and Model Results

The results of the INLA model fit on the chosen predictors by minimizing WAIC are as follows:

Table 1: Poisson Model Fixed Effects

Predictor	Mean	SD
Fed Rate	-0.591	0.134
Unemployment Rate	-0.612	0.098
Home Value	0.208	0.044
Population Work From Home	-0.153	0.046
Mean Travel Time to Work	0.112	0.081
Unpaid Family Workers	-0.034	0.024
Median Household Income	-0.151	0.106
Population Private Health Insurance	0.573	0.213
Population No Health Insurance	0.051	0.025
Opportunity Zone	0.148	0.089
Other Transportation	-0.047	0.033
Ppulation Blue Collar	0.042	0.033
Population w/ Govt Benefits	-0.088	0.047
Population w/ Welfare Benefits	0.239	0.053

Table 2: Poisson Model Hyperparameters and Criteria

Parameter	Mean
Range for Spatial Field	2.01e+04
Precision for Period	0.656
Rho for Period	0.949
Deviance Information Criterion (DIC)	5198.07
Watanabe-Akaike Information Criterion (WAIC)	5205.21
Marginal Log-Likelihood	-2811.69

The magnitude of the fixed effects predictors above can be represents the change in the log of the expected Deal Volume for a one unit change in that predictor.

From the table above, we can see that the strongest positive contributors of Deal Volume is Population Private Health Insurance (.573), Home Value (.208), Population with Welfare Benefits (.239), and Opportunity Zone (.148).

Some possible interpretations for the direction of these coefficeints are as follows:

- **Population Private Health Insurance:** A higher proportion of residents with private health insurance may indicate greater economic stability and disposable income, good indicators for renters, offices, etc.
- **Home Value:** Higher home values reflect a prosperous area, with buyers that are willing to pay high prices
- **Population with Welfare Benefits:** Areas with high welfare benefits might contain undervalued properties in neighborhoods that are on the rise
- **Opportunity Zone:** Designation as an Opportunity Zone provides tax incentives, encouraging investors to engage in more deal transactions within these areas

Conversely, the strongest negative contributors are Unemployment Rate (-.612), Median Household Income (-.151), and Federal Funds Rate (-.591).

Some possible interpretations for the direction of these coefficients are as follows:

- **Unemployment Rate** High unemployment rates may indicate economic instability, lack of disposable income, higher delinquency rates from renters, leasers
- **Median Household Income:** Lower median household incomes may indicate less disposable income for residents to invest in real estate
- **Federal Funds Rate:** Higher interest rates may discourage borrowing and investment in real estate

For the hyperparameters, we see that the range for the spatial field is 20,000 meters. This means that spatial correlation in Deal Volumes extends up to 20 kilometers, indicating that towns within this distance influence each other's deal activities. Much of the towns that lie within 20km of eachother are in the Bay Area, and greater Los Angeles.

The Rho value of .949 is very close to 1, indicating strong temporal correlation in Deal Volumes. This means that Deal Volumes in one period are highly related to those in the subsequent period.

The DIC is 5198.07, the WAIC is 5205.21, and the marginal log-likelihood is -2811.69. These become important when we compare to other types of INLA models.

We can also visualize these results by looking at the posterior distribution of the fixed effects. The posterior distributions for the rest of the fixed effects are contained in the Appendix.

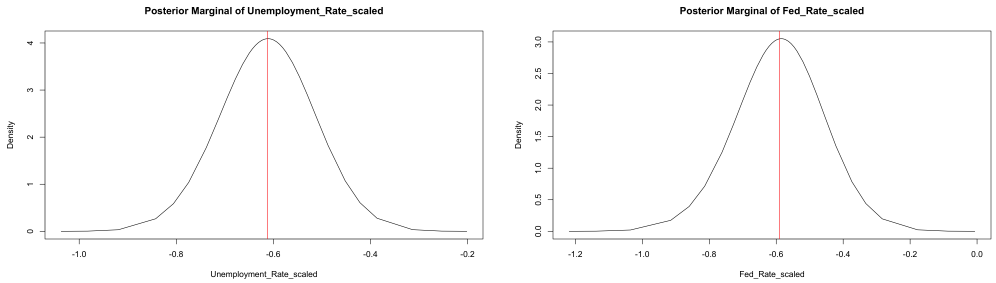

Figure 7: Posterior Distribution of Fixed Effects

Notably, the posterior distribution of the fixed effects appear normally distributed, with a 0 effect contained well outside the credible interval. The same fact is true for Home Value, Median Household Income, the population of, Work from Home, Private Healthcare Holders, Government and Welfare Benefit holders, and areas that are Opportunity Zones .

Since the model is Bayesian, we cannot reject a null-hypothesis for these predictors. However, the credible intervals being far from a zero value means that we can still draw conclusions about the effectiveness of these predictors modeling Deal Volume.

We can also look at the distribution of the spatial and temporal hyperparameters.

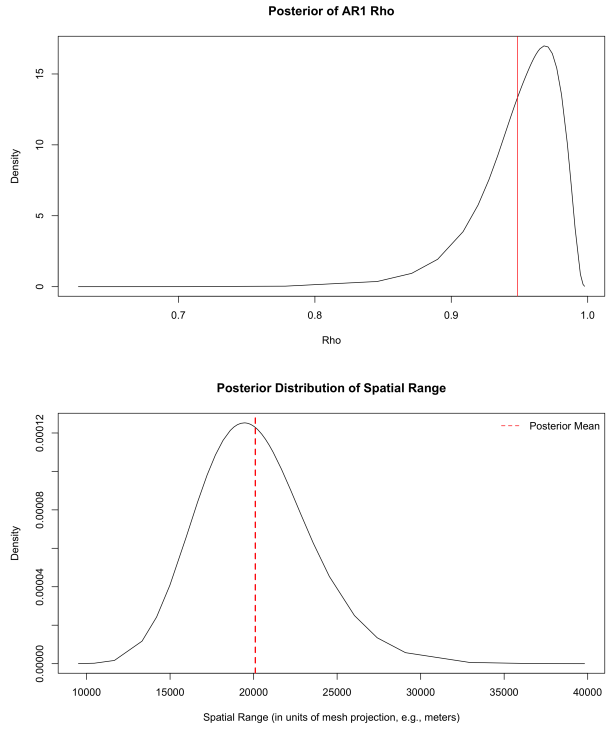

Figure 8: Posterior Distribution of Hyperparameters

Finally, we can examine the mean and standard deviation of the spatial field, which visualizes the estimated spatial effects across the geographic landscape of California towns. Speficaly, this plot reflects how each town's location alone influences its Deal Volume.

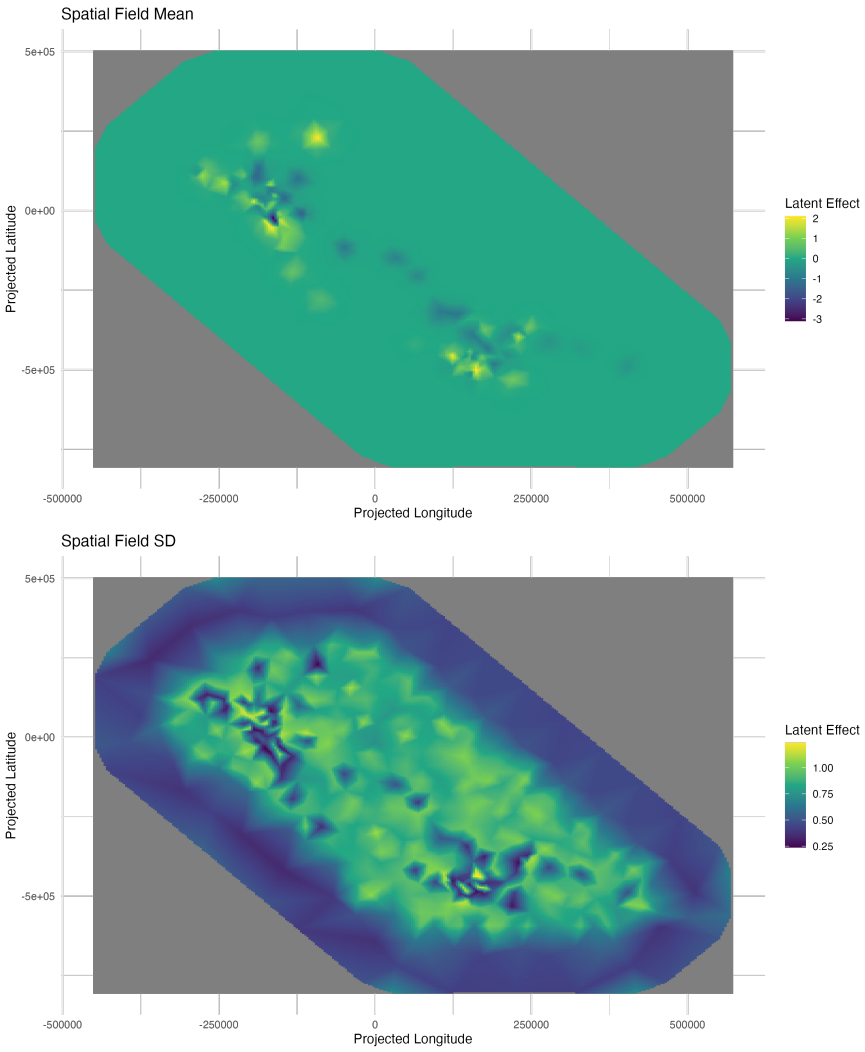

Figure 9: Spatial Effects

Here, we can see how the spatial field forms an outline of California. Within the Spatial Field Mean plot, areas surrounding San Francisco and Los Angeles have a lighter color, meaning that after accounting for all other variables, Deal Volume is higher than the overall average. It is likely that this effect is simply due to the high population density in these areas.

In the Spatial Field Standard Deviation Plot, it is evident that areas with many data points are more certain in their Deal Volume estimates, seen in the dark spots within the visualization indicating lower latent effect (standard deviation). These spots align almost exactly with the bubble plots of towns shown in Section 1.

2.4 Examining Model Performance, Testing Alternative Models

We can visualize the fitted values vs. the observed to evaluate the model's performance:

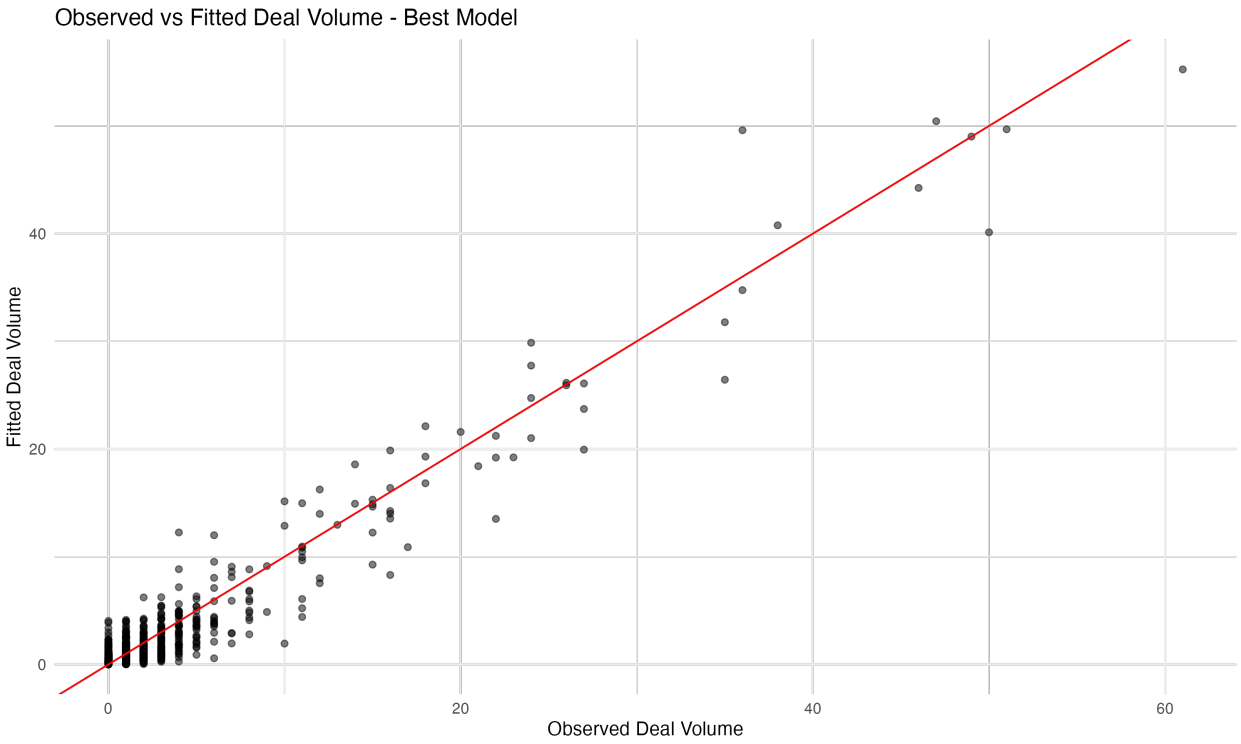

Figure 10: Fitted vs. Observed Values

Based on the plot, we can see the predictive capability of the model is quite adequate. Even at large observed values, the difference between the observed and predicted value is not large, seen in the proximity of the points to the red y = x line.

There is also a clear clustering of points around the 0 observed Deal Volume mark, potentially signaling that our assumption of minimal zero-inflation within the Poisson distribution might be incorrect.

We can also examine how the uncertainty in the mean fitted value changes over time.

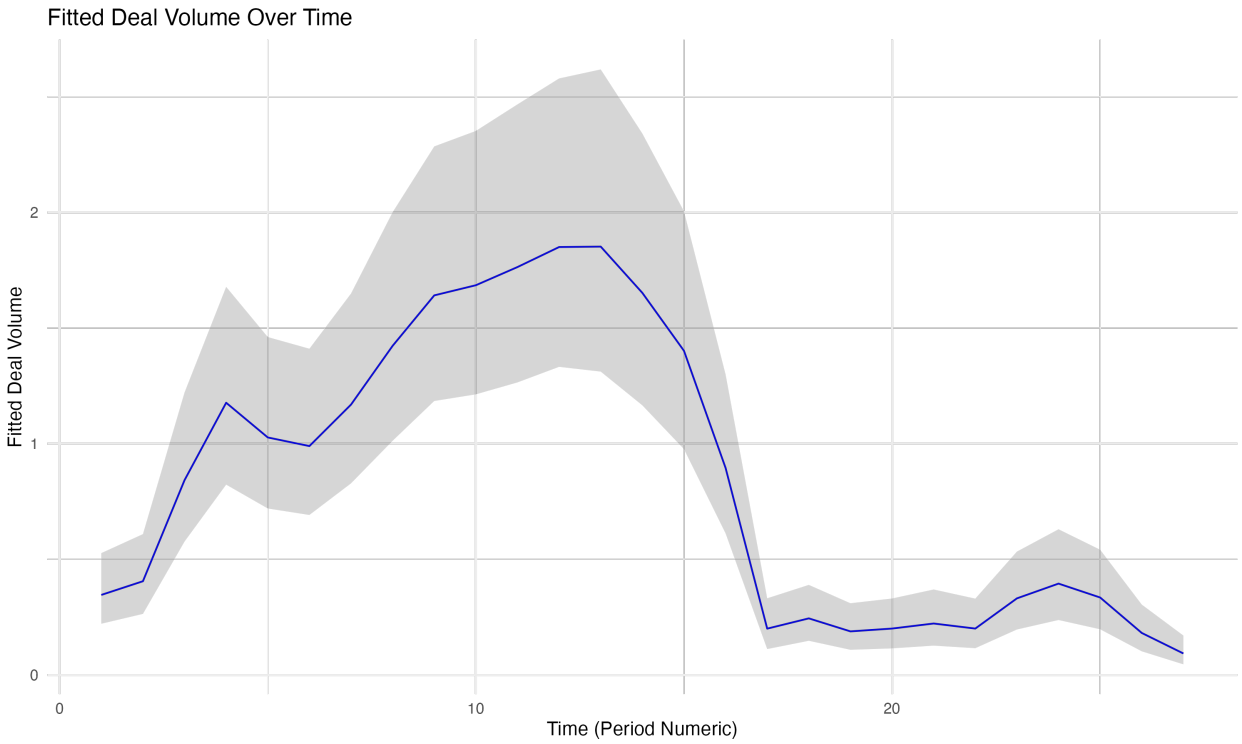

Figure 11: Uncertainty in Fitted Values Over Time

In this plot, we see the uncertainty of the estimate (seen in the gray area surrounding) peaks around time period 15 (equivalent to 2016) at the height of the fitted Deal Volume. This graph is largely similar to Figure 1, the overall deal flow over time. We note positively that as mean Deal Volume increases, the variance does as well - this aligns with the assumption of the Poisson distribution.

Testing a Zero-Inflated Model

Given the abundance of zeroes in the data, a zero-inflated Poisson model is fit within the INLA package. The model can be described as follows,

$$\text{Prob}(y \mid \dots) = p \cdot I[y = 0] + (1 - p) \cdot \text{Poisson}(y \mid y > 0)$$

where p, the probability, is a hyperparameter tuned by a Bayesian prior. We assume that y has some probability of achieving 0, p, and all other values are Poisson distributed.

The model comparison metrics from the fitted zero-inflated model are as follows:

Table 3: Zero-Inflated Model Criteria

Parameter	Value
Deviance Information Criterion (DIC)	6537.91
Deviance Information Criterion (DIC, saturated)	1061.09
Effective number of parameters (DIC)	102.74
Watanabe-Akaike Information Criterion (WAIC)	6550.78
Effective number of parameters (WAIC)	101.56

Marginal Log-Likelihood	-3421.11

Given that the zero-inflated model has a higher DIC and WAIC, as well as more negative log-likelihood, it can be concluded that the original Poisson model is more appropriate for our data.

Testing a Negative Binomial Model

Finally, a negative binomial model was tested, in case there existed overdispersion in the model that would violate assumptions of the Poisson. The results from the model are as follows:

Table 4: Negative Binomial Model Criteria

Parameter	Value
Deviance Information Criterion (DIC)	5542.79
Deviance Information Criterion (DIC, saturated)	1927.04
Effective Number of Parameters (DIC)	123.83
Watanabe-Akaike Information Criterion (WAIC)	5502.35
Effective Number of Parameters (WAIC)	78.44
Marginal Log-Likelihood	-2935.04

The negative binomial model also had a higher DIC and WAIC, and more negative log-likelihood, than the Poisson model. Therefore, the Poisson model is the still the most appropriate for our data.

The following predictive modeling procedure continues with the Poisson model.

2.5 Predictive Modeling

To examine the predictive power of the model, as well as identify potential favorable markets (i.e markets with less actual deals done than predicted), deal data is utilized from June 2023 to January 2024. For all predictor data obtained from the BLS, Federal Reserve, and Zillow, there are available values for this time period. For the census data, the 2024 year is approximated using using the growth rate of the same data point from 2018 to 2023. For those with missing values (perhaps from not being in the census until later), they are assigned the median growth rate. For towns with very large growth rates (usually caused by under reporting in 2018), they are assigned the growth rate of the .75 quartile.

The model is then fit on this mixed of projected and real data, and the predicted Deal Volume is calculated for each town.

The distribution of predicted values only for towns that had deal flow from June 2023 to January 2024 are shown for clarity.

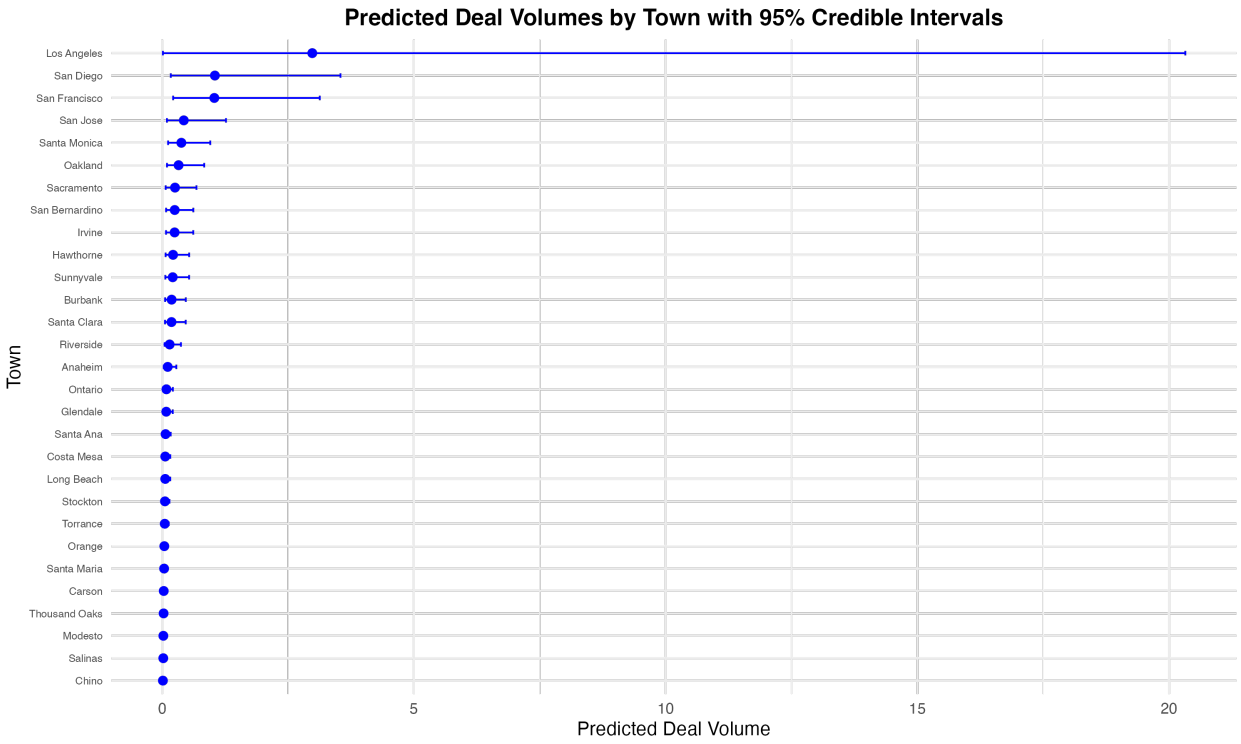

Figure 12: Predicted Deal Volume Distribution

In this plot, towns with the highest predicted Deal Volume also have extremely wide confidence intervals. It appears that variance is greater than the mean in this sense, suggesting that the Poisson distribution might be violated for the predicted values.

We can visualize the predicted values relative to the actual deal flow within the period from June 2023 to January 2024 in the following plot. Here, Predicted_Upper refers to the .975 quartile of predicted values.

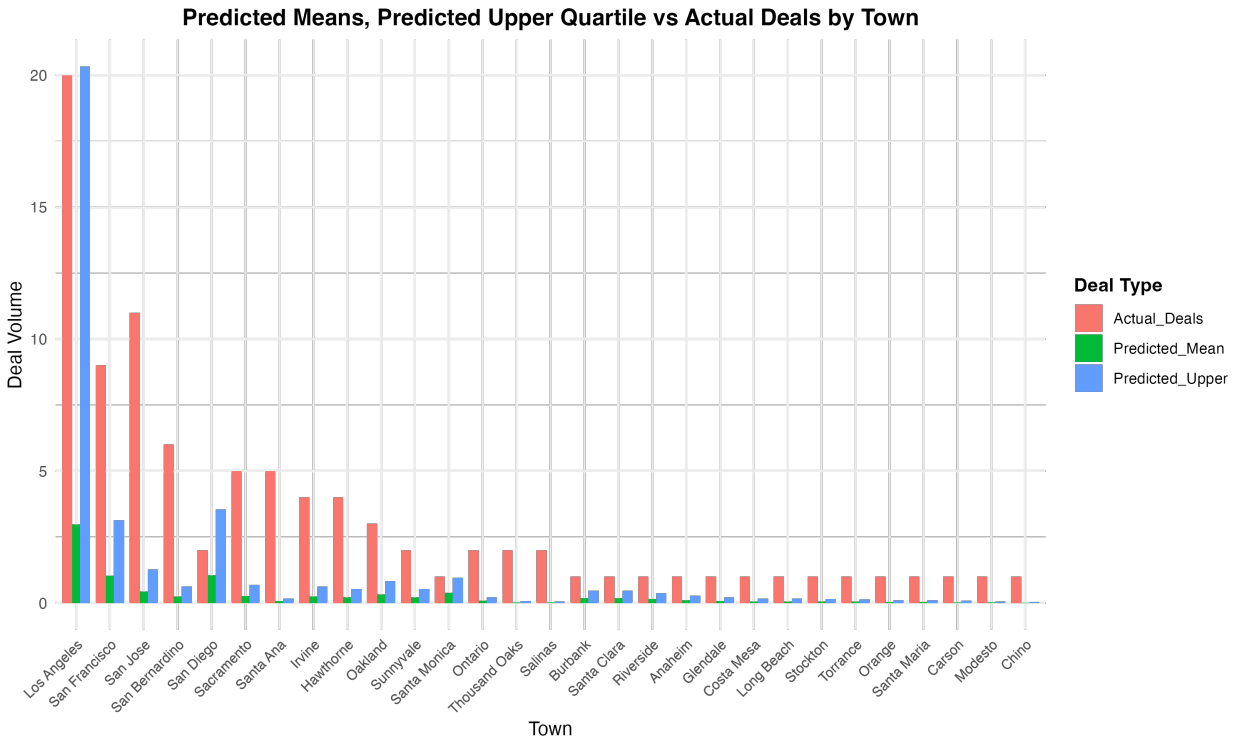

Figure 13: Predicted vs. Actual Deal Volume

For Los Angeles, the .975 quantile actually closely fits the data, suggesting that the model is able to predict the Deal Volume in this area, abeit only at outlier values. The majority of other towns have much greater actual Deal Volume than predicted.

We can visualize this distribution across the map, using the .975 quantile of the predicted values to assess potential favorable markets. The map is shown below:

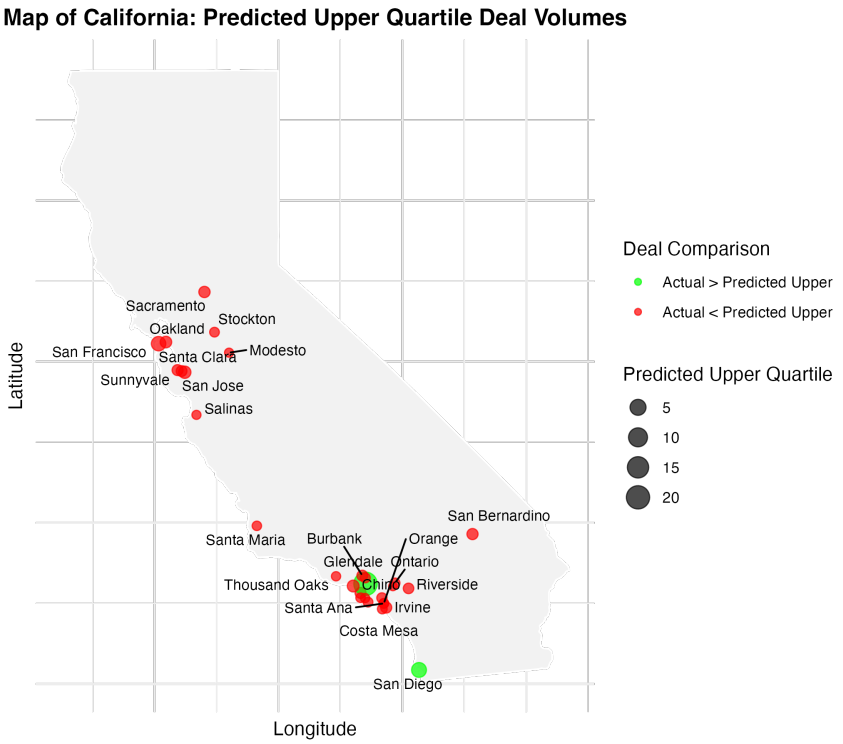

Figure 14: Future Deal Volume by Town

From the plot, we can see that San Diego and Los Angeles are the only towns in which the .975 quantile Deal Volume is greater than the actual. San Diego, in particular, has a mean predicted value quite close to the actual. Perhaps this indicates that the area can be considered favorable.

2.6 Discussion

While the model shows some predictive power, it appears that its primary usage is to identify strong predictors of Deal Volume. The stability of the posterior distributions of the fixed effects leads to reasonable confidence in drawing conclusions about the direction and magnitude of these predictors.

The range of spatial field of 20km makes sense given the clustering of townships around the Bay Area and greater Los Angeles. The strong temporal correlation is also intuitive, as market dislocations and competitive deals are often short-lived and quickly capitalized upon.

However, relative sparsity of deal data across period and across most towns, and the concentration of deal data within large urban areas makes the model less useful for general prediction across townships. The latent spatial effects are likely driven by population density and urbanization, rather than any specific economic or demographic predictor - this is still useful at capturing information within the model, however.

From the model, it is difficult to use prediction to confidently discern any "opportunistic" market without considering outlier predicted values (.975 quantile).

In addition, within the model, there is no distinction made between "opportunistic" and "core" investment, due to the ambiguity of the term. One can imagine the direction of coefficients could be different between a core and opportunistic investor. For instance, a higher Home Value demographic might indicated overvalued properties for an opportunistic investor, but could indicate stable home prices and willing buyers for a core investor.

While interpretation of coefficients can certainly vary, the model still identifies strong broader market signals that are useful in deal flow for both opportunistic and core investors.

3. Conclusion

This analysis identified key macroeconomic and socio-economic predictors that influence Deal Volume in California. The strongest positive contributors to Deal Volume were Population with Private Health Insurance, Home Value, Population with Welfare Benefits, and Opportunity Zone. The strongest negative contributors were Unemployment Rate, Median Household Income, and Federal Funds Rate. The spatio-temporal model identified that Deal Volumes exhibit significant spatial correlation (i.e spillover effects) within a 20-km range. While there is nuance whether these coefficient directions can be associated with *opportunistic* investments specifically, we are still able to identify market signals that drive deal volume in general.

Areas with high urbanization and population density, such as Los Angeles and San Francisco, have higher Deal Volumes. The model also identified strong temporal correlation in Deal Volumes, suggesting that Deal Volumes in one period are highly related to those in the subsequent period.

While the model identifies primarily macroeconomic factors as strong predictors, there is some attributable effect to demographic variables that is useful in identifying favorable markets. The model identified San Diego as a potential market in 2024, albeit with using a wide confidence interval. In future work, it would be useful to expand more predictor types, such as race demographics, crime rates, school quality. In addition, given high sparsity in data across an entire state, one can imagine a more granular analysis, perhaps focusing on Los Angeles itself could be fruitful in identifying more specific predictors of Deal Volume, and boost predictive accuracy. Finally, work that attempts to identify what categorizes as an opportunistic vs. core asset deals could help to further distill specific drivers for each type of investor.

4. References

California Association of REALTORS®. (2018). State of the housing market. In Study of Housing: Insight, Forecast, Trends. https://www.car.org/-/media/CAR/Documents/Industry-360/PDF/Market-Data/2018-State-of-the-Housing-Market-The-SHIFT-Report--Single-Page.pdf

Swensen, D. F. (2000). Pioneering Portfolio Management: an unconventional approach to institutional investment. http://ci.nii.ac.jp/ncid/BA47088905

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019), Spatio-Temporal Statistics with R, Boca Raton, FL: Chapman & Hall/CRC. © 2019 Wikle, Zammit-Mangion, Cressie. https://spacetimewithr.org

5. Appendix

Yearly Deal Volume By Town

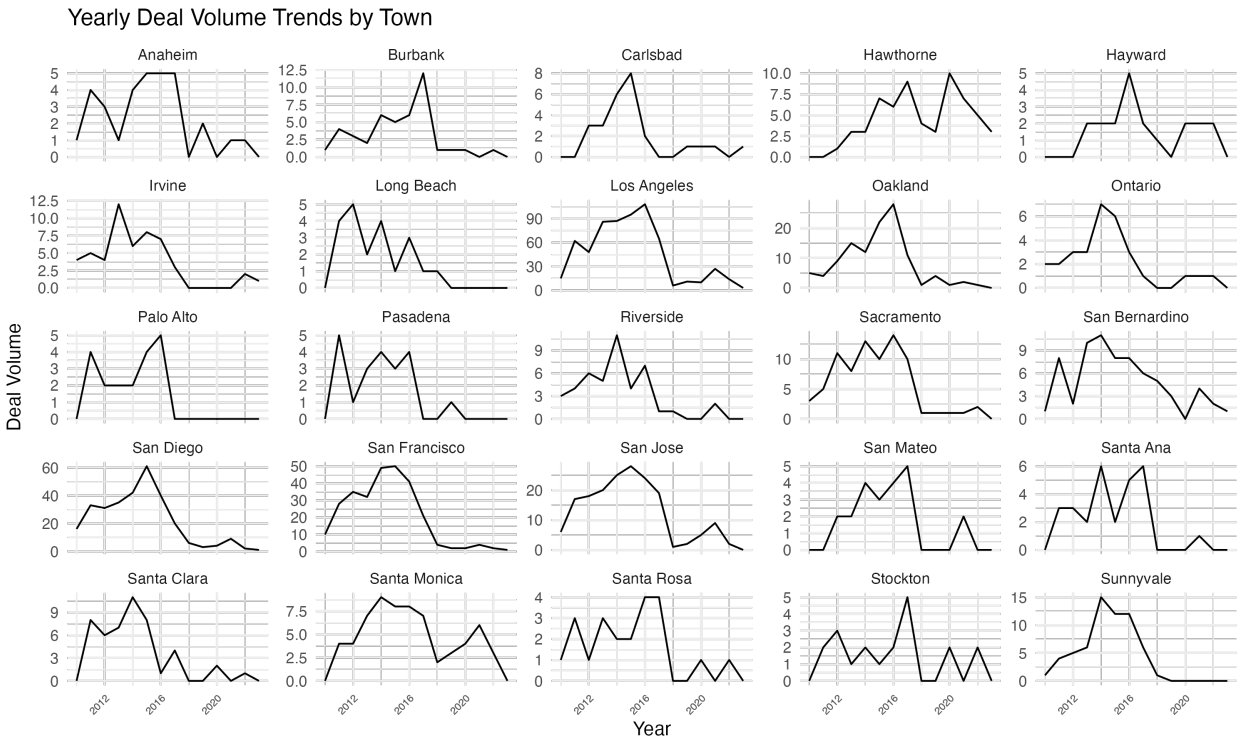

Figure 15: Deal Volume by Town and Year

Strong Predictors by Town (Geographic)

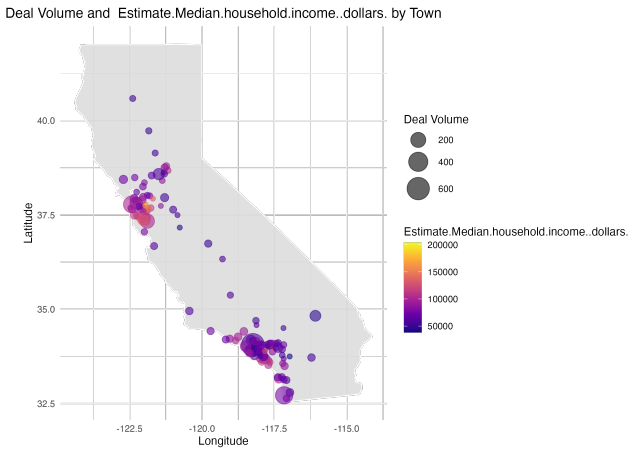

Figure 16: Household Income by Town

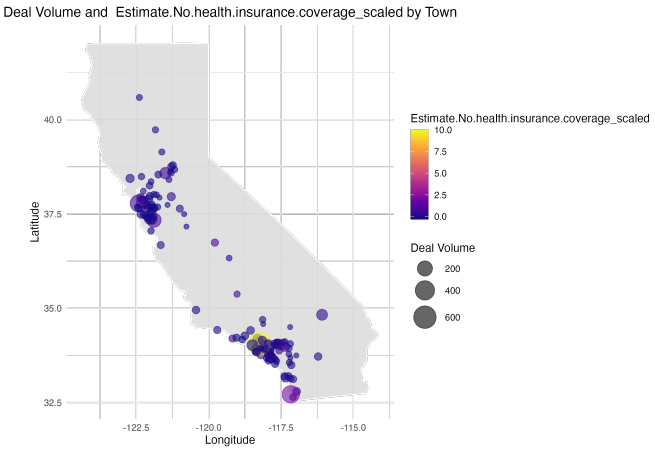

Figure 17: No Health Insurance by Town

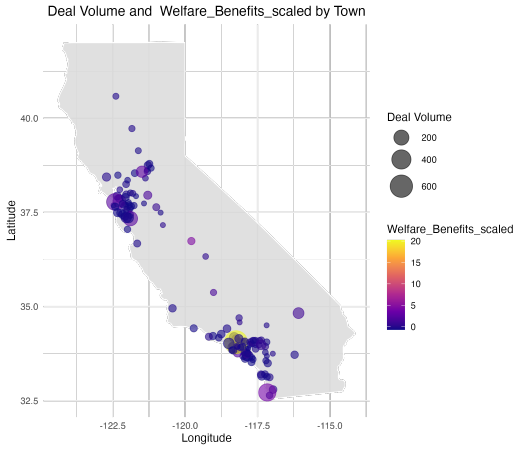

Figure 18: Welfare Benefits by Town

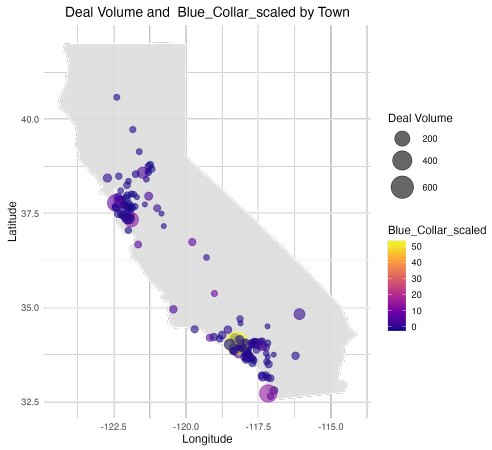

Figure 19: Blue Collar Workforce by Town

2-D Spatial Mesh

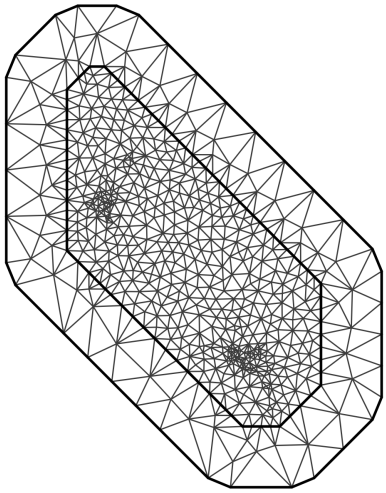

Figure 20: 2-D Spatial Mesh

Posterior Distribution of Fixed Effects

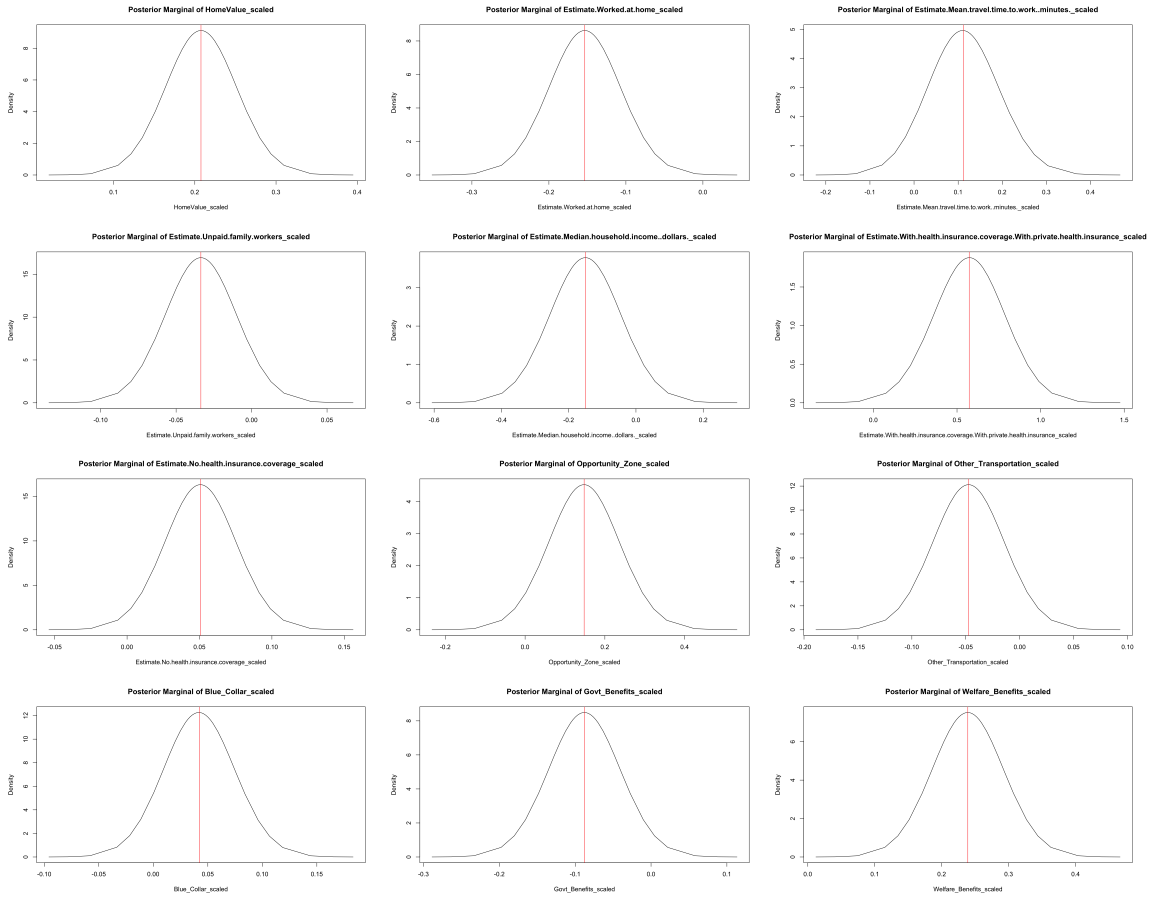

Figure 21: Posterior Distributions of Fixed Effects