California Housing Optimization

Yihan Zhou – yihan_zhou@berkeley.edu

December 18, 2024

Abstract

California's housing affordability crisis poses significant challenges as housing costs surpass income growth. This project provides data-driven recommendations for renters by combining mixed-integer optimization with Simulated Annealing (SA) to identify optimal rental locations. The model minimizes rent-to-income ratios while maximizing access to healthcare, air quality, and safety. SA improves scalability and efficiency, offering near-optimal solutions comparable to exact methods. Results highlight trade-offs between affordability and quality of life, providing actionable insights for renters and policymakers addressing California's housing crisis.

1 Introduction

The goal is to recommend optimal housing locations within the metro areas of California for individuals with specific income levels, balancing affordability and quality of life. To achieve this, I collected data myself and developed a data-driven optimization framework with 2 goals:

1. Minimize: The rent-to-income ratio, ensuring housing is affordable for a given income level.
2. Maximize: Access to key amenities and favorable living conditions, including proximity to healthcare facilities, lower crime rates and better air quality.

2 Optimization Problem Formulation

Decision Variables

x_i: Binary decision variable, $x_i \in \{0, 1\}$ for $i = 1, ..., 30$, where:

$$x_i = \begin{cases} 1 & \text{if region } i \text{ is selected as an optimal housing location} \\ 0 & \text{otherwise.} \end{cases}$$

Here, i refers to a specific metro area in California, as shown in the table below:

Table: California Metro Regions Corresponding to Decision Variables x_i

x_i	**Region**	x_i	**Region**
1	Santa Cruz-Watsonville	16	Sacramento-Roseville-Arden-Arcade
2	San Francisco	17	Riverside-San Bernardino-Ontario
3	San Jose-Sunnyvale-Santa Clara	18	Yolo
4	Santa Maria-Santa Barbara	19	Stockton
5	Salinas	20	Modesto
6	San Diego-Chula Vista-Carlsbad	21	Redding
7	Santa Ana-Anaheim-Irvine	22	Yuba City
8	Napa	23	Fresno
9	Oakland-Fremont	24	Madera
10	Los Angeles-Long Beach-Glendale	25	Chico
11	San Benito County	26	Merced
12	Oxnard-Thousand Oaks-Ventura	27	Hanford-Corcoran
13	Santa Rosa-Petaluma	28	Visalia
14	Vallejo	29	El Centro
15	San Luis Obispo-Paso Robles	30	Bakersfield

Objective Function

The objective function combines affordability and quality of life into a weighted multi-objective optimization:

$$\begin{aligned}\text{Minimize:} \sum_{i=1}^{30} \Bigg(& w_1 \cdot \frac{\text{HourlyWageRequired}_i}{\text{Income}} - w_2 \cdot \text{HealthcareScore}_i \\ & - w_3 \cdot \text{AirQualityScore}_i + w_4 \cdot \text{UnemploymentScore}_i \\ & + w_5 \cdot \text{CrimeScore}_i + w_6 \cdot \text{PopulationScore}_i \Bigg) \cdot x_i\end{aligned}$$

Where:

- w_1, w_2, w_3, w_4, w_5, w_6 : Weights for different factors, reflecting their importance
- $\text{HourlyWageRequired}_i$: Hourly wage required for housing in metro i
- Income: User-defined income
- $\text{HealthcareAccess}_i$: Availability of healthcare facilities in metro i
- AirQualityScore_i: Air quality index in metro i

$$\begin{aligned}\text{AirQualityScore} =& \lambda_1 \cdot \text{Good\_Percentage} + \lambda_2 \cdot \text{Moderate\_Percentage} \\ & + \lambda_3 \cdot \text{Unhealthy\_Sensitive\_Percentage} + \lambda_4 \cdot \text{Unhealthy\_Percentage} \\ & + \lambda_5 \cdot \text{Very\_Unhealthy\_Percentage} + \lambda_6 \cdot \text{Hazardous\_Percentage}\end{aligned}$$

- $\text{UnemploymentScore}_i$: The unemployment rate in metro i
- CrimeScore_i: The crime rate in metro i: $\gamma_1 \cdot \text{Arson Count} + \gamma_2 \cdot \text{Property Crimes Count} + \gamma_3 \cdot \text{Violent Crimes Count}$
- PopulationScore_i: The population per square mile (Land Area) in metro i

Constraints

1. Housing must be affordable: $\text{HourlyWageRequired}_i \leq \alpha \cdot \text{Income}_i, \ \forall i$ where $x_i = 1$, here α represents the maximum allowable rent-to-income ratio
2. Only one metro area can be selected: $\sum_{i=1}^{30} x_i = 1$
3. Non-negativity and binary constraints: $x_i \in \{0, 1\}, \ \forall i \in \{1, ..., 30\}$

3 Data Preparation

I compiled data from authoritative sources to quantify affordability and quality of life across California's metro areas from 2021 to 2024. The dataset includes key metrics such as hourly wages required for housing, adjusted gross income (AGI), unemployment rates, air quality indices, crime statistics (arson, property crimes, and violent crimes), population density, and healthcare facility availability.

The data preparation process involved several steps to ensure the dataset's completeness and consistency for optimization. Missing values in metrics like AGI and crime counts were imputed using linear interpolation or yearly averages within each region, while missing air quality and tax values were filled using regional and temporal averages. Hourly wage data was reshaped from a wide to a long format to facilitate merging, and multi-year data was grouped and aggregated to maintain consistency across years.

Finally, all datasets were integrated based on metro area and year, producing a unified dataset containing all relevant metrics. Key indicators, such as air quality scores and crime rates, were normalized to comparable units to align with the optimization framework.

4 Optimization Models

This project employs Integer Linear Programming (ILP) and Simulated Annealing (SA) to address the optimization problem of selecting an optimal metro area that balances affordability and quality of life.

Method 1: Integer Linear Programming

I firstly used the ILP model, which minimizes a weighted (w_i) combination of six factors: rent-to-income ratio, healthcare facility availability, air quality, unemployment rate, crime levels, and population density. The air quality score is computed as a weighted (λ_i) combination of percentages representing various air quality conditions, while the crime score is determined using a weighted

(γ_i) combination of counts for arson, property crimes, and violent crimes.

The model includes two key constraints: the hourly wage required for housing must not exceed a predefined percentage of the user's income, and only one metro area can be selected. By systematically exploring the feasible solution space, the ILP guarantees an optimal solution, returning the metro area that best balances affordability and quality of life based on the specified weights.

Method 2: Simulated Annealing

Simulated Annealing (SA) is a probabilistic optimization technique inspired by the physical annealing process in metallurgy. Unlike ILP, which guarantees an exact solution, SA explores the solution space by iteratively moving from one solution to a neighboring solution, allowing probabilistic acceptance of worse solutions to escape local optima.

The objective function in SA mirrors the ILP model, combining affordability, healthcare access, air quality, unemployment rate, crime levels, and population density into a weighted formulation. Additionally, **an affordability penalty is added** if the hourly wage required exceeds a predefined percentage of the user's income, ensuring affordability is prioritized.

The algorithm starts with a randomly selected metro area and computes its objective value. At each iteration, a new metro area is chosen as a neighboring solution, and the change in the objective value Δ is evaluated. If the new solution improves the objective $\Delta < 0$, it is accepted. Otherwise, it is accepted probabilistically based on a cooling temperature, which decreases over time, reducing the likelihood of accepting worse solutions. The process continues until the temperature approaches a low threshold or the maximum number of iterations is reached.

SA provides a computationally efficient alternative to ILP, especially for large-scale problems where exact methods may become infeasible. While it does not guarantee an optimal solution, it often finds near-optimal solutions in significantly less time, making it suitable for scenarios requiring scalability and flexibility.

5 Optimization Results

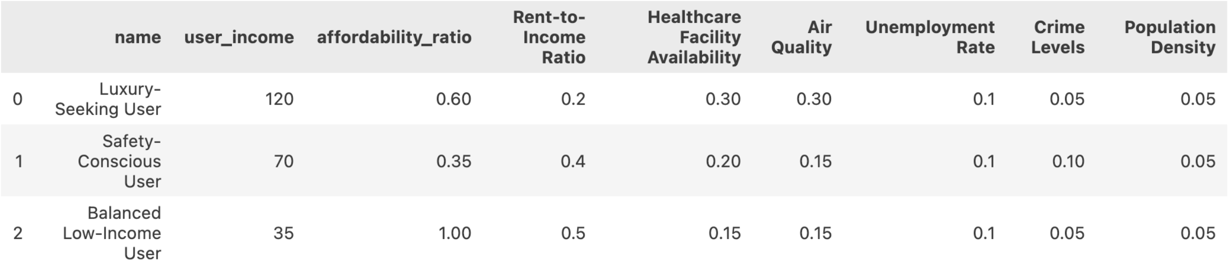
Consistency of two methods

To evaluate the consistency and performance of the proposed optimization methods, I selected a specific parameter configuration: user income of 70 and year 2024. The parameters included an affordability ratio of 0.4, weights for the objective function as $[0.4, 0.2, 0.15, 0.1, 0.1, 0.05]$, air quality weights $[0.6, 0.4, -0.6, -0.8, -1.0, -1.2]$, and crime rate weights $[1, 0.5, 0.7]$.

Both Integer Linear Programming and Simulated Annealing methods were executed under these conditions. The results showed complete agreement between the two approaches, as shown in the figure below. The selected metro area was Chico, and the corresponding objective value was 0.0975. This consistency highlights the reliability of the optimization framework, with ILP providing an exact solution and SA effectively approximating the optimal result within a reasonable computational effort.

Results under different user scenarios

To analyze how income levels and preferences influence housing decisions, various user scenarios were generated. High-income users prioritized factors such as healthcare access and air quality, while affordability constraints were relaxed. For low-income users, the model adjusted for higher rent-to-income ratios to reflect real-world trade-offs.
The parameters for the user cases are summarized below:



	name	user_income	affordability_ratio	Rent-to-Income Ratio	Healthcare Facility Availability	Air Quality	Unemployment Rate	Crime Levels	Population Density
0	Luxury-Seeking User	120	0.60	0.2	0.30	0.30	0.1	0.05	0.05
1	Safety-Conscious User	70	0.35	0.4	0.20	0.15	0.1	0.10	0.05
2	Balanced Low-Income User	35	1.00	0.5	0.15	0.15	0.1	0.05	0.05

Air quality preferences for the users were derived from weighted scores of specific air conditions, as shown below:

Table 1: Air Quality Weights

User Income	Good	Moderate	Unhealthy Sensitive	Unhealthy	Very Unhealthy	Hazardous
120	0.7	0.4	-0.5	-0.7	-1.0	-1.2
70	0.5	0.3	-0.7	-0.8	-1.0	-1.2
35	0.4	0.3	-0.7	-1.0	-1.1	-1.3

Similarly, crime weights were applied across arson, property crimes, and violent crimes:

Table 2: Crime Weights

User Income	Arson	Property Crimes	Violent Crimes
120	0.7	0.5	0.5
70	1.0	0.7	0.6
35	1.4	0.7	0.7

The Integer Linear Programming model was executed for each user case. The results, including the selected metro areas and objective values, are summarized below:

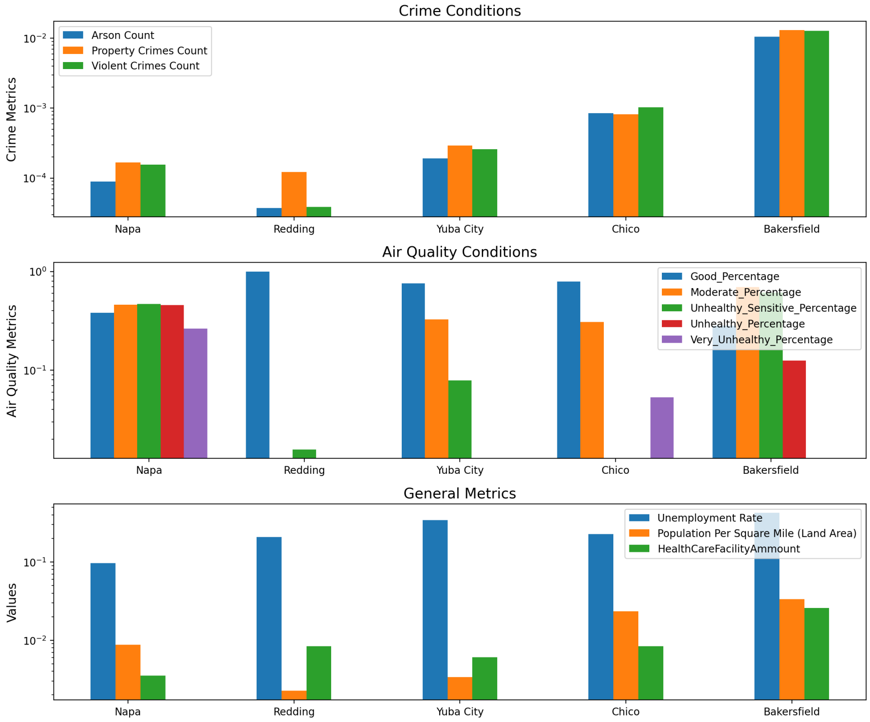
Table 3: Optimization Results for Different User Cases

User Case	Income	Selected Metro	Objective Value
Luxury-Seeking User	120	Redding	-0.141578
Safety-Conscious User	70	Bakersfield	0.206795
Balanced Low-Income User	35	Chico	0.362412

For the *Luxury-Seeking User*, Redding emerged as the optimal metro due to its favorable balance of healthcare access and air quality, despite moderate affordability. The *Safety-Conscious User* was recommended Bakersfield, where crime levels are relatively low. For the *Balanced Low-Income User*, the model identified Chico as the most affordable metro, fulfilling their financial constraints while balancing other factors.

Notably, no feasible solution exists for users with extremely low affordability ratios or strict preferences in areas like air quality or safety. This reflects the limitations of housing availability and affordability in California's current housing market.

To provide further insights into the selected metro areas, key metrics, including crime rates, air quality, and general conditions, were visualized. Figures comparing these metrics for Redding, Bakersfield, and Chico are provided below, with Napa and Yuba City addded as comparison.



The results highlight distinct trade-offs: high-income users prioritize healthcare access and environmental quality over affordability, safety-conscious users accept higher unemployment rates for safer living conditions, and low-income users face significant affordability constraints that limit their housing options.

In conclusion, by integrating Integer Linear Programming and Simulated Annealing, this project can recommend optimal housing locations in California. The optimization successfully balances affordability constraints and quality-of-life factors, including healthcare access, air quality, crime rates, unemployment levels, and population density. Future work can expand the framework by incorporating dynamic housing trends, cost-of-living adjustments, and real-time data integration to improve decision-making.