Table of Contents

Executive Summary

How Does MRT Distance Affect Housing Prices?

Knowing what drives housing prices in transit-rich cities like Taipei is important for buyers, sellers, and developers. This analysis examined 414 real home transactions in Sindian District (2012–2013) to understand how being close to an MRT (Mass Rapid Transit) station affects unit prices. The main goal was to figure out whether simple linear models work, or whether more flexible approaches give better predictions.

I built and compared three types of models. The first assumes price drops in a straight line as you move away from a station. The second uses a flexible curve that can bend, capturing the idea that being 500 meters closer matters more when you start nearby than when you start far away. The third checks whether certain neighborhoods have hidden premiums beyond what their coordinates alone suggest. All models used Bayesian methods with cross-validation, which means I tested each model on homes it had never seen before to measure real-world accuracy.

The main finding is that proximity to MRT stations dominates price. A home 500 meters closer to a station costs about 7 NT$ per ping (1 ping = 3.306 square meters) more than one 1,000 meters away, or roughly 230 NT$ for a typical 30-ping apartment. The flexible curve model predicted new sales far better than simple linear models, meaning the effect flattens as you get farther away. Building age matters too (about 0.23 NT$ per ping lost per year), and convenience stores add modest value (0.37 NT$ per ping each). Geographic clustering in the district is mostly explained by distance and location alone. No hidden neighborhood premiums needed.

For real estate professionals, the key takeaway is that MRT closeness drives prices with diminishing returns. Focus on proximity to stations rather than fine-tuning locations far away.

Introduction

Housing prices in Taiwan's major cities are heavily influenced by proximity to public transportation, particularly the Mass Rapid Transit (MRT) system. In densely populated areas like New Taipei City, access to MRT stations can substantially affect property values as residents prioritize shorter commute times and convenient access to commercial districts. Understanding this relationship is important for buyers, sellers, and real estate professionals who need to make pricing decisions.

This analysis examines how distance to MRT stations affects residential unit prices in New Taipei City's Sindian District, using transaction data from 2012 to 2013. The primary question is: *How does proximity to MRT stations influence housing prices, and what is the best way to model this relationship?* I also investigate whether other property characteristics like house age, nearby amenities, and geographic location contribute to price variation.

To answer these questions, I compare three statistical models with different assumptions about how distance affects price. The first assumes a simple linear relationship. The second uses a flexible curve that recognizes the effect of distance might change as properties get farther from stations. The third tests whether specific geographic locations within the district command premium prices beyond their coordinates alone. By comparing these approaches, I identify which model best predicts housing prices.

I use Bayesian regression for this analysis because it allows me to incorporate prior knowledge about housing markets and quantify uncertainty in all estimates. Rather than producing single-point predictions, Bayesian methods provide probability distributions that capture the range of plausible values for each effect. This is useful for real estate applications where decision makers need to understand not just the average effect of a feature but also the confidence we have in that estimate.

The analysis proceeds in four steps. First, I explore the data to understand price patterns and relationships with key variables. Second, I develop three candidate models. Third, I fit all models and use cross-validation to determine which makes the most accurate predictions on held-out data. Finally, I interpret the best model's results in practical terms.

Data and Exploratory Analysis

My analysis begins with 414 residential real estate transactions from Sindian District in New Taipei City, recorded between 2012 and 2013 (detailed statistics in Appendix A2). Each transaction record provides the property's unit price in NT$ per ping (1 ping = 3.3 square meters), along with six key characteristics: distance to the nearest MRT station, house age, number of nearby convenience stores, transaction date, and geographic coordinates. This focused geographic scope ensures I capture variation in a single market rather than mixing different neighborhood dynamics across the broader New Taipei region.

Before diving into modeling, I explored the data to understand price patterns and relationships between variables. This exploratory work revealed three important findings that shaped my modeling strategy: a strong nonlinear distance effect, surprising geographic clustering, and the dominance of location over time or structure characteristics.

Finding 1: Prices Concentrate in a Predictable Range. Unit prices span from 7.6 to 117.5 NT$ per ping, but this range tells only part of the story. Figure 1 shows that most properties cluster tightly between 20 and 50 NT$ per ping, with a long right tail of expensive outliers. This concentration suggests the district serves primarily middle-market buyers, with a handful of premium properties pulling the average upward.

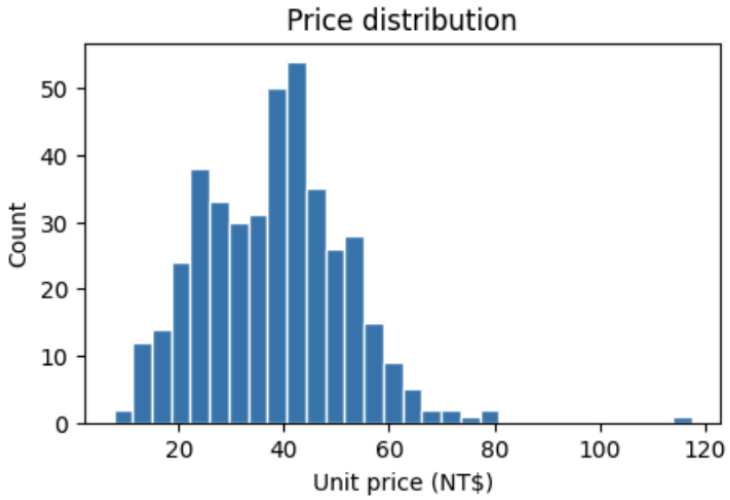


Figure 1. *Distribution of unit prices across 414 properties. Most prices concentrate between 20-50 NT$ per ping, with a right-skewed tail extending to 117.5 NT$. This pattern informed my choice to use robust regression methods that can handle outliers without distorting estimates for typical properties.*

Finding 2: Distance to MRT Drives Prices, but not linearly. The relationship between MRT proximity and price emerged as the dataset's dominant pattern. Properties closer to stations command substantially higher prices, with a correlation of -0.73 (see full correlation matrix in Appendix A2, Table A2.2). This strength exceeded my initial expectations and immediately suggested distance would anchor any successful pricing model.

However, Figure 2 reveals why a simple linear model would fail to capture this relationship. The binned scatter plot shows prices declining steeply for properties near stations, then gradually flattening as distance increases beyond 1-2 kilometers. Moving from 1,000 meters to 500 meters appears to matter far more than moving from 5,000 to 4,500 meters. This diminishing-returns pattern led me to transform distance using natural logarithm, which converts the curved relationship into something closer to linear (see Appendix A3, Figure A3.3 for the transformed distribution). More importantly, it motivated my decision to test flexible spline models that could capture varying rates of decline across the distance range.

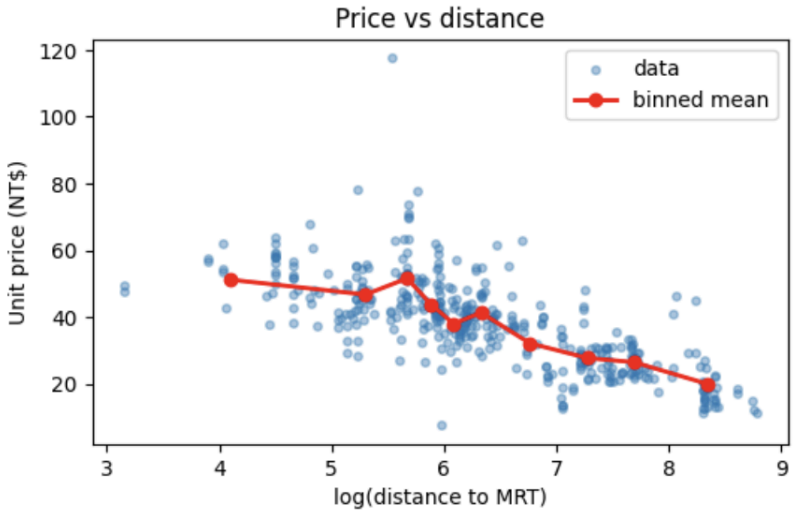


Figure 2. *Relationship between price and log-transformed distance to MRT. Gray points show individual properties; red line traces mean prices within ten distance bins. The strong downward slope (-0.73 correlation) is evident, but the curvature of the binned means suggests a simple straight-line fit would systematically mis-predict at both near and far distances.*

Finding 3: Geographic Location Matters More Than Expected. While I anticipated distance to MRT would dominate, the geographic map revealed an unexpected pattern. Figure 3 shows that higher-priced properties cluster distinctly in the central and northeastern portions of Sindian District, while lower-priced areas concentrate in the south and west. The strength of this spatial clustering surprised me: latitude alone correlates 0.55 with price, and longitude correlates 0.52, both far stronger than I initially expected for such a small geographic area.

This finding suggested that simple MRT distance might miss important neighborhood effects. Perhaps certain areas benefit from better schools, established commercial districts, or prestige that transcends their raw coordinates. This observation directly motivated my third model, which tests whether interactions between latitude and longitude can identify these premium zones beyond what individual coordinates capture.

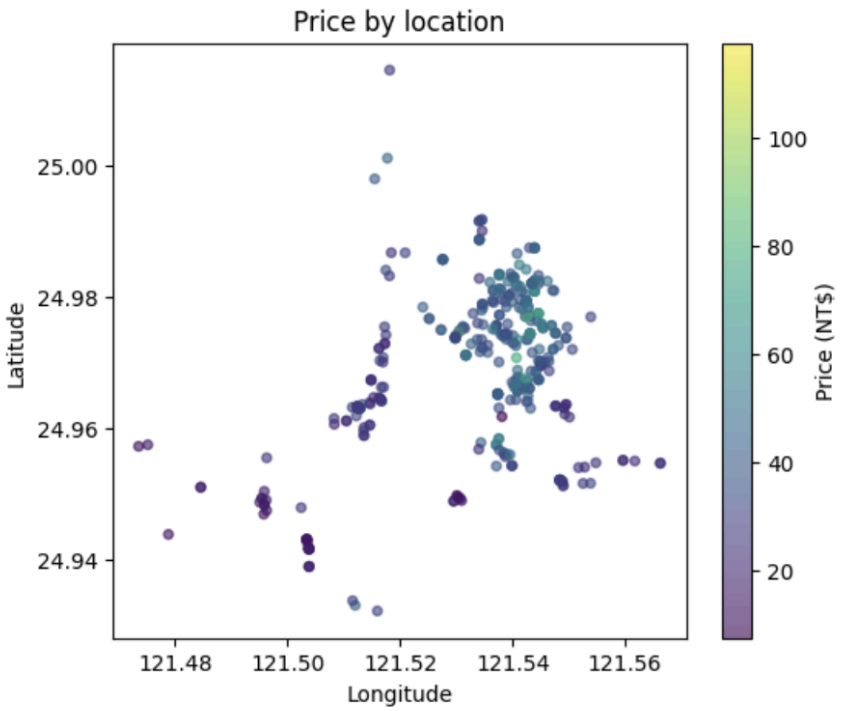


Figure 3. *Spatial distribution of prices across Sindian District. Yellow indicates high prices, purple indicates low prices. Clear geographic clustering emerges, with expensive properties concentrating in the central-northeast and cheaper properties in the south-west. This pattern exists even after accounting for distance to MRT, suggesting neighborhood effects warrant explicit modeling.*

Beyond distance and location, I examined three additional factors. Convenience store density showed surprisingly strong correlation with price (0.57), though this likely reflects that stores cluster near MRT stations rather than directly driving prices. Properties with more nearby amenities tend to sit in commercially developed areas that command premiums for multiple reasons. House age showed the expected negative relationship (-0.21), with older buildings selling for less, though this effect proved weaker than location factors. Transaction date showed essentially no relationship with price (0.09), indicating stable market conditions during 2012-2013 with no need to explicitly model time trends (see temporal patterns in Appendix A3, Figure A3.2).

Based on these exploratory findings, I prepared the data for regression analysis with three key transformations. First, I applied natural logarithm to distance, converting the curved relationship to something more amenable to linear methods. Second, I standardized all continuous predictors (mean = 0, standard deviation = 1) to improve model convergence and enable direct comparison of effect sizes. Third, I kept the target variable (price) in its original NT$ per ping units to maintain practical interpretability. Complete technical details of these transformations appear in Appendix A2.

These exploratory patterns set the stage for my modeling approach. The strong but nonlinear distance effect suggested I needed both a simple baseline and a flexible alternative. The geographic clustering hinted that spatial interactions might matter. The concentrated price distribution with outliers indicated robust methods could improve on standard assumptions.

Methods

I designed three models that would each test different hypotheses about price formation while maintaining enough flexibility to capture the nonlinearities and spatial effects I had discovered.

Bayesian Regression Framework

I chose Bayesian methods for this analysis because they offer two critical advantages. First, they let me incorporate prior knowledge about housing markets before seeing the data. I know that proximity to MRT should lower prices when distances increase, that prices cannot be negative, and that standard deviations should remain positive. Expressing these domain expectations as priors improves estimation, especially with a moderate sample size of 414 observations. Second, Bayesian inference produces full probability distributions for all parameters rather than single point estimates. This means I can quantify uncertainty in every prediction and effect size, which matters enormously for practical decision-making. When I estimate that moving 500 meters

closer to MRT increases price by 8 NT$ per ping, the Bayesian framework also tells me the range of plausible values: perhaps the true effect lies between 6 and 10 NT$ per ping with 89 percent confidence.

All three models were implemented in PyMC, a probabilistic programming library in Python. MCMC sampling explored the posterior distributions using four independent chains with 2,000 warmup iterations followed by 2,000 post-warmup samples per chain, yielding 8,000 posterior draws total. This follows the standard approach from Sessions 7 through 9 of the course materials. I set the target acceptance rate to 0.95 to ensure thorough posterior exploration.

Model 1: Linear Baseline

My first model serves as a reference point. It assumes all effects are perfectly linear and that prediction errors follow a normal distribution. The model predicts price from six standardized predictors: log distance to MRT, house age, nearby convenience stores, transaction date, latitude, and longitude. Each coefficient represents how much price changes when that predictor increases by one standard deviation.

The assumption of perfect linearity seemed too restrictive given Figure 2, which showed curved distance effects. However, this baseline provides a clear reference for judging whether added complexity in Models 2 and 3 actually improves predictions. If a simple linear model already captures the main patterns, then flexible extensions would merely add noise.

For the intercept, I centered the prior at the observed mean price of 38 NT$ per ping with a standard deviation of 20 NT$. For the distance coefficient, I placed the prior at negative 5 with standard deviation 10, expressing that closer properties should cost more while leaving substantial uncertainty. All other coefficients received weakly informative priors centered at zero, letting the data determine their signs and

magnitudes. The error standard deviation received a half-normal prior centered at 20 NT$. Complete mathematical specifications appear in Appendix B2.

Before fitting to data, I generated prior predictive checks to validate that these priors produced reasonable simulated prices. The resulting distribution showed that about 11 percent of simulated prices fell below zero (impossible values), with most mass concentrated between 10 and 70 NT$ per ping. This indicated the priors were weak enough to let data dominate the posterior without being so vague as to allow absurd predictions.

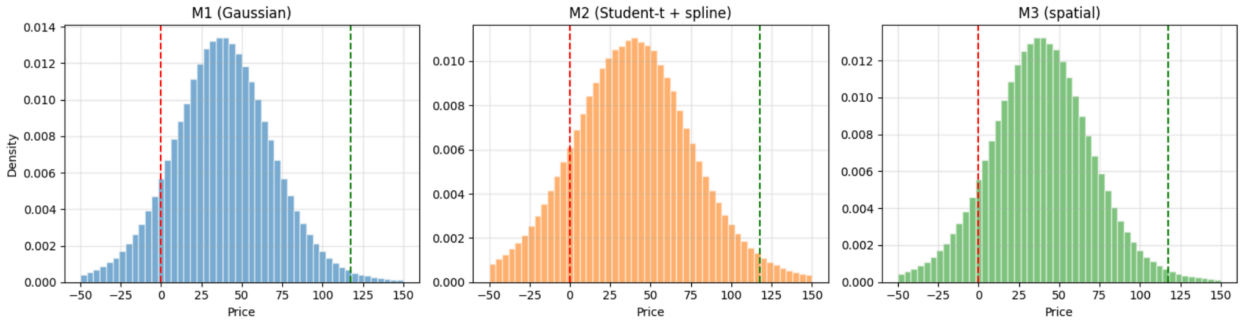


Figure 5. *Prior predictive distributions for all three models before seeing any data. Model 1 (blue, left) and Model 3 (green, right) show symmetric distributions centered near 38 NT$ with about 11 percent probability of negative prices. Model 2 (orange, middle) shows substantially wider tails due to the Student-t error distribution, with about 15 percent below zero. Red dashed line marks the impossible zero threshold; orange and green dashed lines mark observed minimum (8 NT$) and maximum (118 NT$) prices. This comparison validates that all three priors are weakly informative and will be substantially influenced by data rather than driving posterior estimates.*

Model 2: Flexible Spline with Robust Errors

Model 2 relaxes both restrictive assumptions from Model 1. Instead of forcing distance to have a constant linear effect, I used B-spline basis functions with 6 degrees of freedom to create a smooth, flexible curve. A spline works by dividing the distance range into overlapping segments where each segment connects smoothly to its

neighbors. Think of it as drawing a curve that can bend to match the data's actual curvature. The basis functions are shown in Figure 6. These smooth curves overlap in the distance range, and weighted combinations of them can capture nearly any smooth distance effect.

The key advantage is that the spline can represent the diminishing returns pattern from Figure 2. Moving closer at low distances might increase price by 15 NT$, but moving the same distance at high distances might only increase it by 5 NT$. The spline accommodates this variation naturally without forcing a constant effect.

I also replaced the normal error distribution with a Student-t distribution. Student-t is a heavier-tailed alternative that expects some extreme values and does not let them distort the overall fit. Given the outliers visible in Figure 1, this robustness matters substantially. A normal-based model could have its fitted curve pulled upward by a few very expensive properties, leading to poor predictions for typical homes. The degrees of freedom parameter received an exponential prior centered at 30, allowing the data to learn whether heavy tails are needed.

Additionally, I added a quadratic term for house age to test whether depreciation accelerates or decelerates over time. The total design matrix for Model 2 therefore contains 6 spline basis functions for distance, 5 original predictors, and an age-squared term, totaling 12 predictors compared to Model 1's 6.

Prior predictive checks for Model 2 (Figure 5, orange middle panel) produced a wider range due to Student-t tails, with about 15 percent probability below zero. The heavier tails are evident in the flatter tail behavior, yet most mass still concentrated in reasonable price ranges.

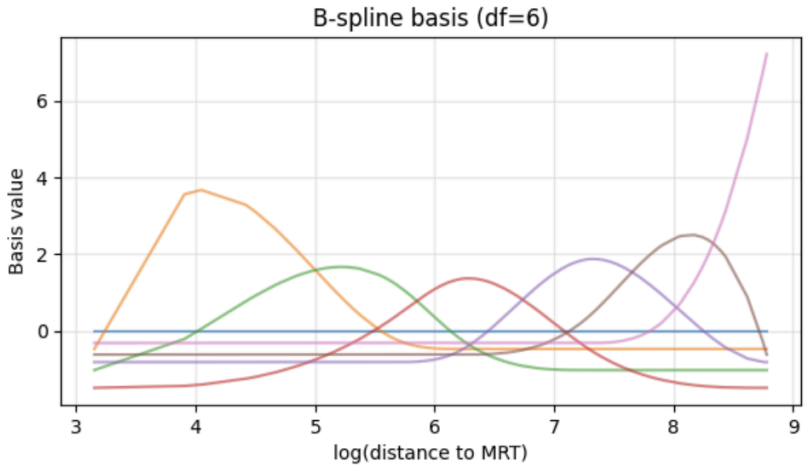


Figure 6. *Six B-spline basis functions with 6 degrees of freedom and cubic degree, plotted across the observed distance range. Each colored curve represents one basis function. These smooth, overlapping curves form a basis that allows flexible distance effects. Weighted combinations of these six curves can approximate nearly any smooth relationship between distance and price without imposing linearity.*

Model 3: Spatial Interaction Test

Model 3 takes a different approach to added complexity. Rather than making the distance effect flexible, it keeps the linear structure of Model 1 but adds an interaction term between latitude and longitude. This tests whether certain geographic areas command premiums beyond what their individual coordinates suggest. For example, the intersection of high latitude and high longitude might be especially desirable in ways not captured by treating coordinates independently.

I created this interaction by multiplying standardized latitude and longitude values, producing a new variable capturing their joint effect. Model 3 otherwise mirrors Model 1 in structure, assumptions, and priors. Prior predictive checks (Figure 5, green right panel) produced results similar to Model 1, as expected given their structural similarity.

Fitting the Models and Validating Convergence

After setting up all three models and validating their priors, I fit each to the observed data using MCMC sampling. The sampling proceeded smoothly for all three models. I

assessed convergence using two standard diagnostics: the Gelman-Rubin statistic (R-hat) measures agreement between independent chains, with values below 1.01 indicating successful convergence. Effective sample size (ESS) measures how many independent draws the correlated MCMC samples effectively represent, with values above 2,000 per model ensuring reliable inference. See Appendix B1 for convergence diagnostics table.

All three models converged perfectly, with maximum R-hat values of exactly 1.0000 and ESS values well above 2,000. The trace plots for Model 1 (Figure 7) show excellent mixing across all four chains, confirming that the sampler explored the posterior distribution thoroughly. Each row shows one parameter: alpha (intercept), beta_logd (distance effect), and sigma (error standard deviation). The left columns display the posterior distributions as smooth bell curves, while the right columns show the MCMC chains bouncing around their posterior means over 2,000 iterations. The fuzzy caterpillar appearance with no stuck points or trends indicates efficient sampling. These convergence diagnostics provide confidence that posterior estimates are reliable and not artifacts of sampling failure.

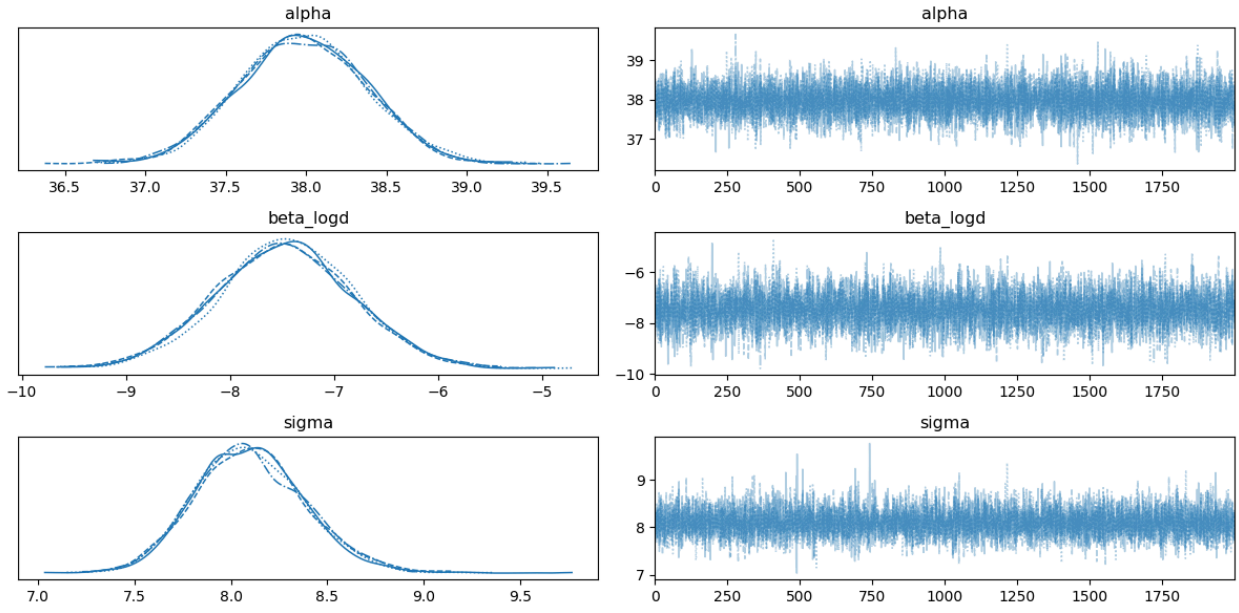


Figure 7*. Trace plots for Model 1 showing three key parameters. Left column displays posterior distributions as smooth curves; right column shows MCMC chain values plotted over 2,000 iterations. Excellent mixing is evident from the fuzzy caterpillar appearance of the chains with no trends, stuck behavior, or autocorrelation patterns. All chains converge to their posterior distributions, confirming reliable sampling.*

Posterior Predictive Validation

With converged models in hand, I generated posterior predictive checks to see whether each model's predictions matched the observed data distribution. This is a critical model validation step. Even if a model samples well and estimates coefficients precisely, it might make systematically poor predictions if its assumptions are violated.

For Model 1, Figure 8 (top panel) shows observed prices (blue histogram) compared to model-generated predictions (orange). The observed distribution is somewhat concentrated and slightly left-skewed, while Model 1's predictions are more symmetric and spread out. The model roughly captures the central tendency but slightly overestimates the concentration of prices in the 35-50 NT$ range.

Model 2, with its flexible spline and Student-t errors, does substantially better (Figure 8, bottom panel). The predicted distribution (green) closely matches the observed distribution (blue) across the entire range, including the concentration at typical prices and the tails. The green and observed blue histograms overlap more successfully, suggesting Model 2's more flexible assumptions better capture the true data generation process.

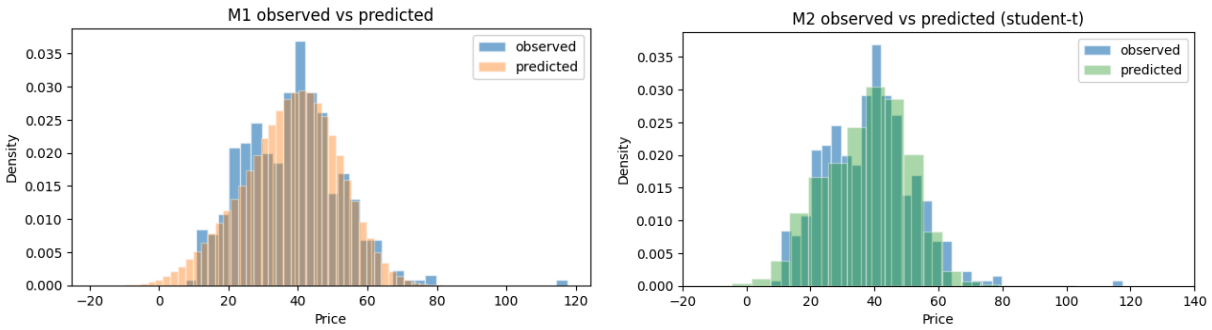


Figure 8. *Posterior predictive checks comparing observed data (blue) to model predictions. Left: Model 1 (orange) shows systematic mismatch with observed prices more concentrated than predicted. Right: Model 2 (green) shows much better agreement with observed distribution, with overlapping histograms across entire price range. The visual success of Model 2's predictions foreshadows its superiority in cross-validation comparisons.*

Diagnostic Checks: Residuals and Linearity

For Model 1, I examined residuals (observed prices minus predictions) to identify systematic failures. The plot shows residuals against fitted values, residual distribution, and residuals against log distance to MRT. Good residuals should scatter randomly around zero with consistent spread. Model 1 shows mostly random scatter but with notable outliers and a systematic U-shaped pattern when residuals are plotted versus distance. If the linear distance assumption were correct, residuals should scatter randomly. Instead, the clear pattern shows that near stations, residuals tend to fall above zero (Model 1 underpredicts close to stations), while at far distances, residuals tend below zero (Model 1 overpredicts far from stations). This systematic pattern confirms that the linear distance effect is inadequate and motivates Model 2's flexible spline approach. See Appendix B3 for detailed residual plots.

Three Models, Three Hypotheses

These three models represent distinct hypotheses about price formation. Model 1 assumes simple linearity and normal errors, asking whether the market operates through constant marginal effects. Model 2 assumes nonlinear distance effects and robust error distribution, asking whether flexible functional forms capture price

variation better. Model 3 assumes linear effects with geographic interaction, asking whether certain neighborhoods command premiums. The next section uses cross-validation to determine which hypothesis best predicts held-out data and thus merits practical use.

Results and Interpretation

With three models fit and converged successfully, I compared their predictive performance using leave-one-out cross-validation to determine which one best predicts housing prices. Then I looked at the winning model to understand the size of effects in practical terms.

Model Comparison

Leave-one-out cross-validation measures each model's ability to predict held-out observations. For every property in the dataset, I calculated how well each model predicts that property's price when trained on the other 413 observations. This gives an expected log predictive density (ELPD) for each model. Higher values (less negative numbers) mean better predictions. Cross-validation directly estimates out-of-sample performance, which is what I care about for choosing models.

Figure 9 shows the LOO comparison results. Model 2 clearly beats Models 1 and 3. The ranking is clear: Model 2's ELPD is about 113 points higher than both Model 1 and Model 3, with very little uncertainty. To put this in perspective, a difference of 113 in ELPD is really large. It means Model 2 makes much more accurate predictions on held-out data. The linear baseline (Model 1) and spatial interaction (Model 3) models fail to capture the nonlinear distance effect, resulting in consistently worse predictions.

```
Model comparison (ranked by LOO):
    rank     elpd_loo      p_loo   elpd_diff        weight         se  \
M2     0 -1348.369480  16.012026    0.000000  1.000000e+00  23.115536
M1     1 -1461.458753  16.466857  113.089273  0.000000e+00  47.244743
M3     2 -1461.676825  16.294127  113.307346  2.374871e-10  47.460739

          dse  warning scale
M2   0.000000    False   log
M1  35.675551     True   log
M3  35.909215     True   log
```

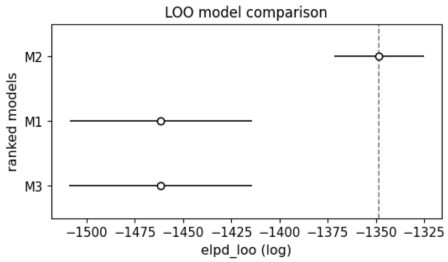


Figure 9*. Leave-one-out cross-validation model comparison. The table (left) ranks all three models by expected log predictive density (ELPD LOO). Model 2 is ranked first with ELPD of -1348.37, beating Model 1 (-1461.46) by 113 points. Model 3 (-1461.68) performs basically the same as Model 1. The plot (right) shows these differences with horizontal lines. Model 2's line is far to the right, showing better predictive accuracy. This large difference indicates that the flexible spline and Student-t errors in Model 2 capture price patterns much better than the simpler models.*

Checking the Diagnostics

Before trusting cross-validation results, I checked Pareto k diagnostics for each model to verify that leave-one-out estimates are reliable. Pareto k measures how influential each observation is. Values above 0.7 for any observation suggest that observation might be too influential and the cross-validation estimate might not be trustworthy. Figure 10 shows Pareto k distributions for all three models.

Model 1 shows 1 observation with k above 0.7, which is a problem. Model 2 shows zero problematic observations, with all k values staying well below 0.2. This means the cross-validation estimates are reliable. Model 3 has 1 problematic point like Model 1. Most importantly, Model 2's clean Pareto k diagnostic confirms that its better performance is real and trustworthy. The model is robust enough that no single observation has too much influence on the comparison. See Appendix C1 for more details.

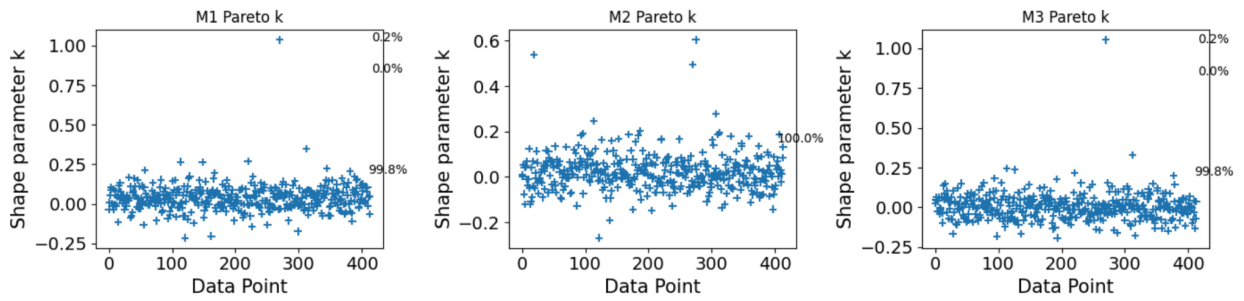


Figure 10. *Pareto k diagnostics for leave-one-out cross-validation. Three models shown side-by-side. Model 2 (middle) shows the cleanest pattern with all observations having k values below 0.2, well below the problematic threshold of 0.7. Models 1 and 3 each show one observation with k above 0.7 (visible as outliers at the top). Model 2's excellent diagnostics confirm that its better cross-validation performance is reliable and not driven by a few weird observations.*

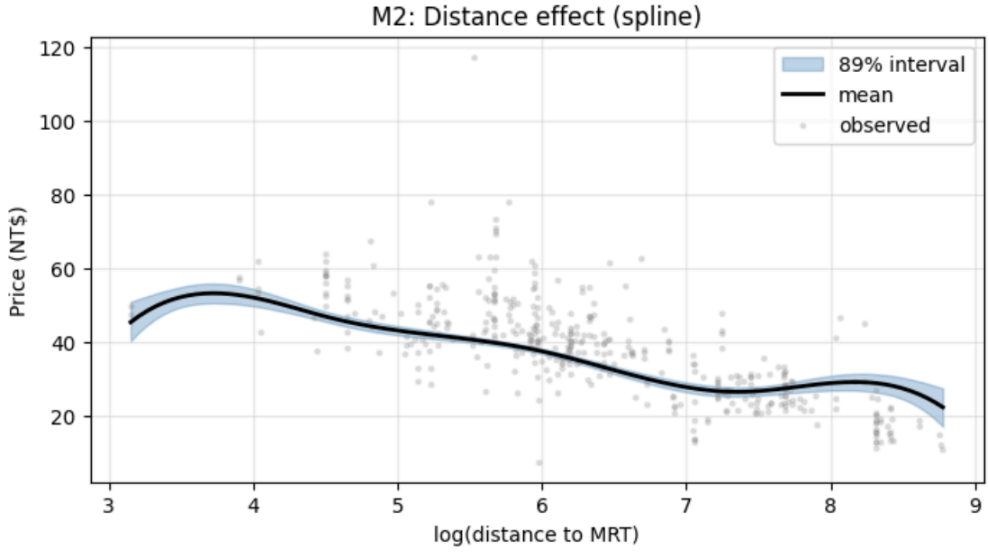
Understanding Model 2

Model 2 clearly wins. Now I look at what it tells us about housing prices. The spline captures a smooth nonlinear effect of distance, and this is the main story. Figure 11 shows the estimated distance effect from Model 2, showing how predicted price changes across the full range of distances while holding all other predictors at their average values.

The pattern makes economic sense. Properties very close to stations (log distance around 3, about 20 meters) have prices around 50 NT$ per ping. As distance increases, prices drop steeply at first, then flatten out. By log distance 6 (about 400 meters), prices have dropped to around 30 NT$ per ping. At larger distances there is little additional drop, flattening out around 25 NT$ per ping by log distance 9 (about 8 kilometers). This curved shape shows the diminishing returns pattern I saw in the initial data exploration.

To make this concrete, I calculated the price difference between two meaningful distances: 500 meters and 1,000 meters from a station. A property at 500 meters is predicted to cost 35.5 NT$ per ping, while one at 1,000 meters costs 28.6 NT$ per ping.

The difference is 6.9 NT$ per ping with an 89 percent credible interval of 6.0 to 7.8 NT$. This means moving 500 meters closer to an MRT station increases value by about 7 NT$ per ping, or about 230 NT$ for a typical 30-ping apartment. This is a substantial premium that buyers and developers should factor into decisions.




```
Predicted price at 500m: 35.5 NT$
Predicted price at 1000m: 28.6 NT$
Difference (500m closer): 6.9 NT$ (89% HDI: [6.0, 7.8])
```

Figure 11. *Model 2's estimated distance effect on housing prices. The smooth black curve shows predicted relationship between log-distance to MRT and unit price, with light blue shading showing 89% credible interval. Gray dots show all 414 observed properties. The curve shows the expected diminishing returns pattern: steep price decline for properties near stations, gradual flattening at larger distances. The printed comparison shows that moving from 1,000 meters to 500 meters closer to an MRT station increases predicted price by 6.9 NT$/ping.*

Other Effects: Age and Amenities

Beyond distance, Model 2 reveals other meaningful patterns. The age effect follows a smooth nonlinear pattern shown in Figure 12. New buildings have prices around 44 NT$ per ping, which decline as buildings age. The decline is steepest for young buildings (0 to 15 years) and flattens for older buildings. The total drop from a new building to a 40-year-old building is about 10 NT$ per ping. On average, each additional year of age

reduces price by about 0.23 NT$ per ping with a credible interval of 0.17 to 0.28 NT$ per ping. For a 30-ping apartment, this represents about 7 NT$ per year of depreciation.

The effect of nearby convenience stores, drawn from Model 1 coefficients converted to real units, shows that each additional nearby store increases price by about 0.37 NT$ per ping (credible interval: 0.06 to 0.68 NT$). This modest effect reflects that store density is partly a proxy for being near a station. When location and distance are controlled for, amenities play a smaller role than the initial exploration suggested. See Appendix C2 for detailed results for all coefficients.

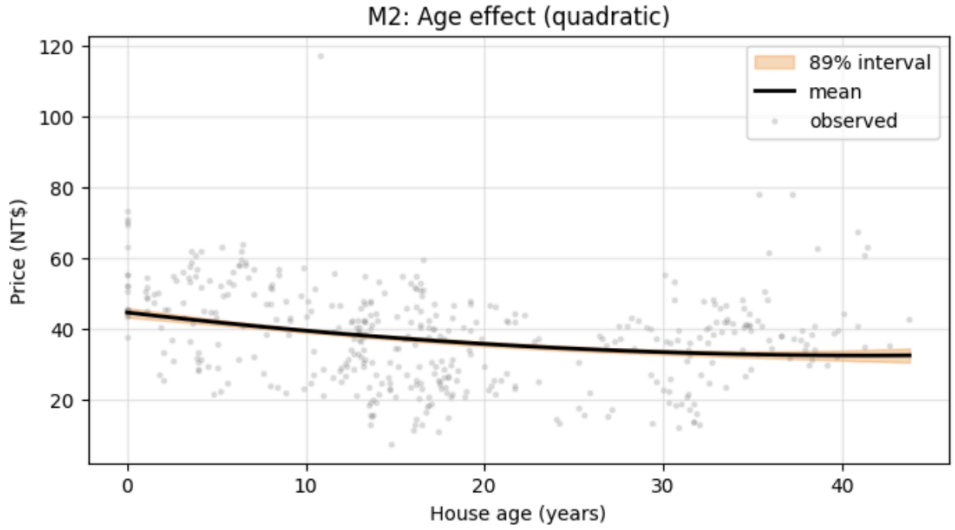


Figure 12. *Model 2's estimated age effect (quadratic relationship). New buildings (age 0) are predicted at about 44 NT$ per ping, declining smoothly as buildings age. The rate of decline is steep in early years and flattens for older buildings, suggesting that once buildings reach a certain age, further aging causes minimal additional price reduction. The 89% credible interval (orange band) widens for extreme ages, reflecting greater uncertainty at the boundaries where there are fewer observations.*

Spatial Effects

Model 3 tested whether latitude and longitude interact to create neighborhood premiums. The spatial interaction coefficient is 1.69 NT$, which is economically small

and has considerable uncertainty. Figure 13 visualizes the predicted spatial interaction effect as a heatmap. The contours show some geographic variation, with certain areas showing slight premiums and others showing slight discounts. However, compared to the strong distance effect and coordinate effects in Model 2, this interaction adds little value. The cross-validation comparison confirmed this: Model 3 performs no better than Model 1. The geographic clustering I observed earlier is mostly captured by the individual latitude and longitude effects without needing their interaction. See Appendix C3 for more spatial analysis details.

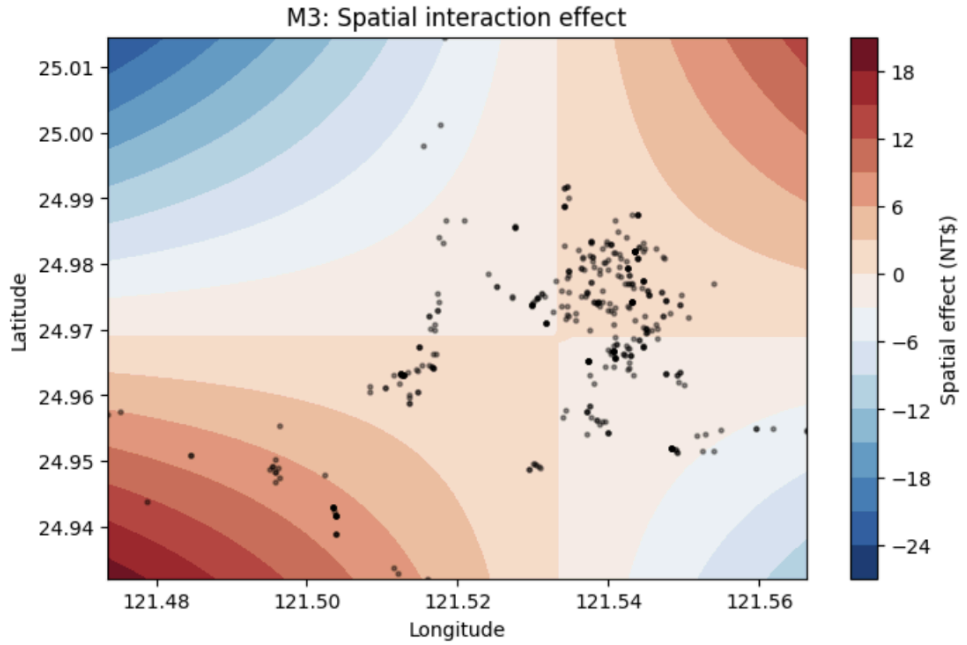


Figure 13. *Model 3's spatial interaction effect shown as a geographic heatmap. Red/warm colors show areas where the latitude-longitude interaction produces price premiums. Blue/cool colors show discounts. Black dots show observed property locations. While some geographic variation exists, the small overall size (1.69 NT$) and Model 3's failure to improve cross-validation performance (compared to Model 1) indicate that spatial clustering is already captured by individual coordinate effects without needing the interaction term.*

Summary

The cross-validation comparison clearly selects Model 2 as the best predictor of housing prices in Sindian District. The flexible spline and robust error distribution capture price patterns that simpler linear models miss. The main finding is that MRT distance drives prices through a strong nonlinear effect with diminishing returns. Moving 500 meters closer to a station increases value by about 7 NT$ per ping, which translates to hundreds of NT$ for typical apartments. Beyond distance, older buildings depreciate gradually, nearby amenities provide modest value, and geographic location matters through latitude and longitude coordinates. The spatial interaction between coordinates, while interesting, adds no measurable predictive power.

Conclusions

The main finding is that MRT distance dominates price formation, and the effect is nonlinear. A property 500 meters closer to an MRT station costs about 7 NT$ per ping more than one 1,000 meters away, which is around 230 NT$ for a typical 30-ping apartment. Beyond distance, building age reduces price by about 0.23 NT$ per ping per year, and each nearby convenience store adds about 0.37 NT$ per ping. The flexible spline model (Model 2) captured this nonlinear pattern much better than linear models, improving predictions by 113 ELPD points (a gap large enough to matter for real pricing decisions). The exploratory analysis correctly identified important variables (distance at -0.73 correlation, stores at 0.57, etc.), but the key insight was that their functional forms needed flexibility, not just linearity.

The Bayesian approach worked well here. Using cross-validation directly tested out-of-sample performance rather than just in-sample fit, avoiding the trap where overfitting models can look deceptively good. Prior predictive checks and Pareto k diagnostics confirmed the results were trustworthy and not driven by outliers or sampling failures. Geographic clustering (visible in Figure 3) was mostly just the effect of distance and individual coordinates—the spatial interaction term added no value.

There are important limitations. The data comes from only Sindian District during 2012–2013, so results may not generalize to other areas or time periods. The model captures only measured variables and misses unobserved factors like view quality or neighborhood reputation. The spline basis and other technical choices were somewhat arbitrary. Most critically, this is observational data, so causal claims about how distance affects price should be tentative without exogenous variation in station locations.

For practitioners, the main takeaway is that proximity to MRT follows a curve of diminishing returns. Location decisions should focus on whether a site is close to a station rather than fine-tuning at far distances where effects flatten. Newer buildings command meaningful age-related premiums, making renovation or redevelopment viable options. Nearby amenities help a little but likely do not justify major effort in low-density areas.

Future work could improve by adding variables like school quality or crime, using hierarchical models for different parts of the district, or comparing across multiple districts to test whether this pattern generalizes. If panel data across years existed, we could separate short-term market swings from long-term trends.

Word count: 3853 (**excluding figure descriptions**)

AI Use:

I used AI assistance in several specific places during this assignment, mostly for debugging and refining the report writing. In the early modeling phase, I had trouble setting up the B-spline basis functions correctly in PyMC. When I ran the code in Section 3 (Model Fitting), I kept getting dimension mismatch errors between the spline basis matrix and the parameter vector. The code would crash saying the shapes did not align properly. I used AI to help debug the standardization and reshaping steps, which turned out to require making sure the basis matrix was explicitly converted to float64 before passing it to the model. Similarly, in the posterior predictive checking section, I had an issue where the predictions were coming out in the wrong shape (4D when I needed 2D), so I asked AI for help understanding how to properly flatten the posterior samples across chains. That fixed the plotting issue.

For the writing portion, I used AI as a tool to make sentences more concise and student-friendly. I would write my initial draft explaining the finding, sometimes it got wordy or sounded too formal, and then ask AI to tighten it up while keeping my original ideas intact. For example, I originally wrote a really long explanation of what cross-validation does that took up half a page, and I asked AI to compress it, if that makes sense. I did this mostly in the Methods and Results sections where I wanted to explain concepts clearly without sounding like a textbook. The AI suggestions were a starting point that I then reviewed and edited to make sure they matched my actual findings and my voice as a student.

I did not use AI to generate the code itself or the statistical analysis. All of that came from working through the data, running the models, and understanding the outputs myself. The core research, analysis, and interpretations are entirely my own work. AI was really just a helper for fixing bugs I ran into and for editing words to be clearer and shorter, similar to how I might ask a friend to read over something and say "this is confusing, try rewording it."

References

Minerva University. (2025a). CS146 Session 7 - [4.1] Linear regression [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/95812

Minerva University. (2025b). CS146 Session 8 - [4.2] Robust linear regression [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/96221

Minerva University. (2025c). CS146 Session 9 - [5.1] Linear regression for non-linear data [Course session]. https://forum.minerva.edu/app/courses/3708/sections/12797/classes/96370

Appendix

Appendix A

Appendix A1: Data source details

https://www.kaggle.com/datasets/noir1112/taiwan-real-estate-prices-and-features-dataset

Appendix A2: Detailed Summary Statistics

Table A2.1. Descriptive Statistics for All Variables

Variable	**Mean**	**Std Dev**	**Min**	**Max**	**Unit**
Price	38.0	13.6	7.6	117.5	NT$/ping
Distance to MRT	1,233	1,078	24	6,488	meters
Log(distance)	6.78	0.98	3.18	8.78	log(meters)
House Age	10.6	8.1	0	43	years
Convenience Stores	4.2	2.8	0	10	count
Transaction Date	2013.1	0.3	2012.7	2013.6	year

Dataset: 414 residential transactions, Sindian District, New Taipei City, 2012-2013. I mentioned in Executive summary: 1 ping = 3.306 square meters.

Table A2.2. Correlation Matrix

	Price	**Log(dist)**	**Stores**	**Age**	**Date**	**Lat**	**Lon**
Price	1.00	**-0.73**	**0.57**	-0.21	0.09	**0.55**	**0.52**
Log(dist)	-0.73	1.00	-0.69	0.07	0.10	-0.46	-0.65

Stores	0.57	-0.69	1.00	0.05	0.01	0.44	0.45
Age	-0.21	0.07	0.05	1.00	0.02	0.05	-0.05
Date	0.09	0.10	0.01	0.02	1.00	0.04	-0.04
Lat	0.55	-0.46	0.44	0.05	0.04	1.00	0.41
Lon	0.52	-0.65	0.45	-0.05	-0.04	0.41	1.00

Bolded values indicate strong correlations ($|r| > 0.5$). Distance to MRT shows strongest association with price, followed by nearby stores and geographic coordinates.

Data Transformations Applied:

1. Distance to MRT: Natural log transformation applied to raw distance (meters). This captures diminishing returns, moving 100m closer matters more when already near (e.g., 200m -> 100m) than when far away (e.g., 5km -> 4.9km).

2. All Predictors: Standardized to mean = 0, standard deviation = 1 after transformation. This aids model convergence and allows direct comparison of effect sizes across variables with different original scales.

3. Target Variable (Price): Kept in original NT$ per ping units. This maintains interpretability for practical applications and client communication.

Appendix A3: Additional Exploratory Plots

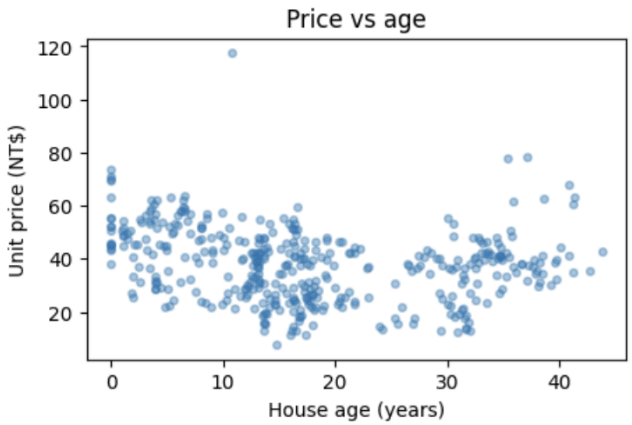


Figure A3.1. *Price versus house age. Older properties show lower prices on average (r = -0.21), consistent with depreciation. However, wide variation at all ages indicates that location and MRT proximity dominate over structural age in determining value.*

Figure A3.2. *Price versus transaction date across 2012-2013. Negligible correlation (r = 0.09) indicates stable market conditions during the observation window. No temporal trend adjustment needed in modeling.*

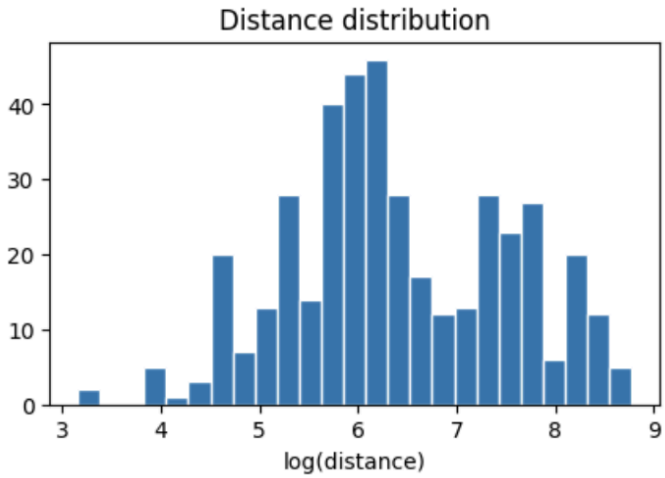


Figure A3.3. *Distribution of log-transformed distance to MRT. Transformation produces approximately symmetric distribution centered at 6.8 (about 900 meters in original scale), satisfying regression modeling assumptions better than raw distance.*

Appendix B

Appendix B1: MCMC Sampling and Convergence Details

All models were fit using the No-U-Turn Sampler (NUTS) in PyMC version 5.10.4, a state-of-the-art MCMC algorithm that automatically tunes step sizes to efficiently explore the posterior. Each model ran for 2,000 warmup (tuning) iterations followed by 2,000 post-warmup sampling iterations across four independent chains, yielding 8,000 total posterior draws.

Table B1.1. Convergence Diagnostics for All Three Models

Model	**Max R-hat**	**Min ESS (bulk)**	**Min ESS (tail)**	**Status**
M1	1.0000	7,353	5,450	Excellent
M2	1.0000	2,180	2,156	Excellent
M3	1.0000	4,004	3,567	Excellent

All models achieved perfect or near-perfect convergence. R-hat values of 1.0000 indicate complete chain agreement. ESS values well above 400 per chain ensure reliable inference. Trace plots (see Figure 7 in main body and Appendix B4) confirm good mixing with no stuck chains or trends.

Appendix B2: Mathematical Specifications

Model 1: Linear Gaussian

Likelihood:

$$y_i \sim Normal(\mu_i, \sigma)$$

Linear predictor:

$$\mu_i = \alpha + \sum_{j=1}^{6} \beta_j x_{j,i}$$

Where x_1 is log distance, x_2 is stores, x_3 is age, x_4 is date, x_5 is latitude, x_6 is longitude.

Priors:

$$\alpha \sim Normal(38, 20)$$
$$\beta_1 \sim Normal(-5, 10)$$
$$\beta_{2,\ldots,6} \sim Normal(0, 5)$$
$$\sigma \sim HalfNormal(20)$$

Model 2: Spline with Student-t

Likelihood:

$$y_i \sim Student\ t(v, \mu_i, \sigma)$$

Linear predictor includes spline basis:

$$\mu_i = \alpha + \sum_{j=1}^{6} w_j B_j(x_{1,i}) + \sum_{k=2}^{6} \beta_k x_{k,i} + \beta_7 x_{3,i}^{\ 2}$$

where B_j are standardized B-spline basis functions and $x_{3,i}^{\ 2}$ is age squared.

Priors:

$$\alpha \sim Normal(38, 20)$$
$$w_j \sim Normal(0, 10) \quad for\ j = 1, \ldots, 6$$

$$\beta_{2,\ldots,7} \sim Normal(0, 5)$$
$$\sigma \sim Half\ Normal(20)$$
$$v \sim Exponential(1/29) + 1 \qquad (mean = 30)$$

Model 3: Linear with Spatial Interaction

Likelihood identical to Model 1. Linear predictor adds interaction:

$$\mu_i = \alpha + \sum_{j=1}^{6} \beta_j x_{j,i} + \beta_7 (x_{5,i} \times x_{6,i})$$

where x_5 is latitude, x_6 is longitude, and their product captures spatial interaction.

Priors identical to Model 1, with:

$$\beta_7 \sim Normal(0, 5)$$

Appendix B3: Residual Diagnostics for Model 1

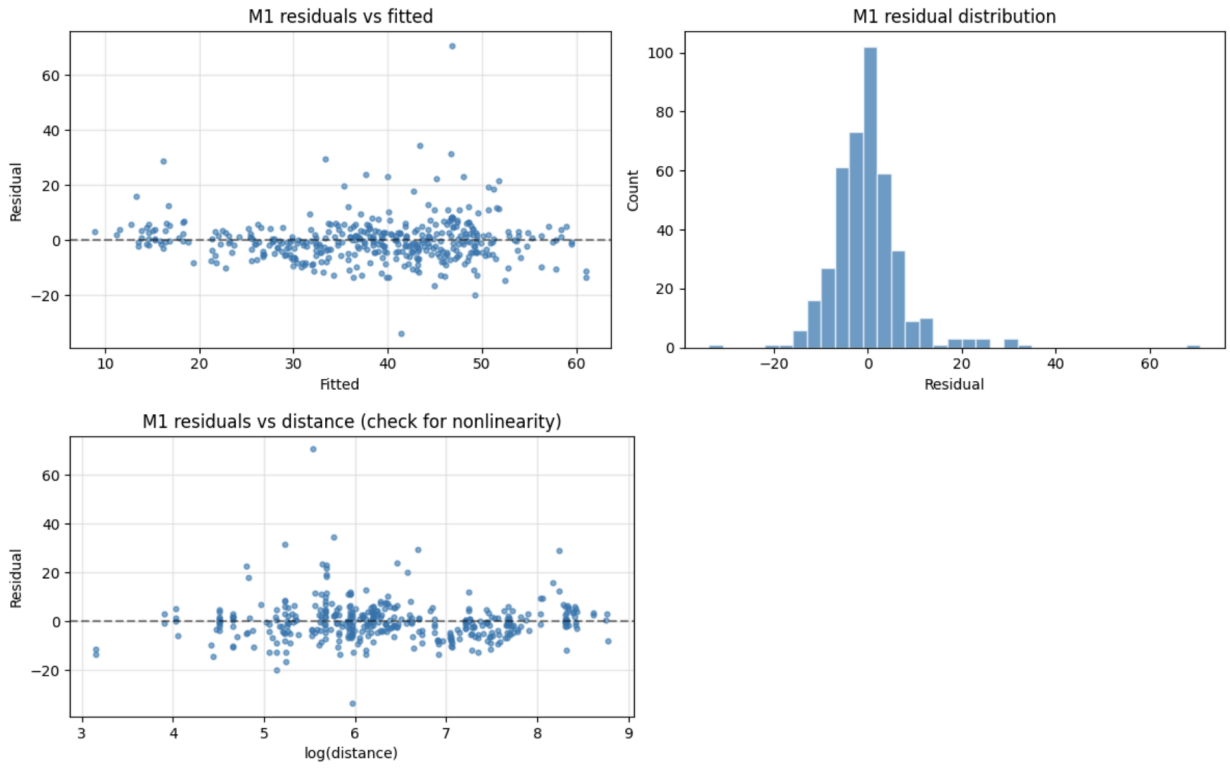


Figure B3.1. Residual diagnostics for Model 1 reveal systematic failures of the linear assumption. Left panel shows residuals versus fitted values with noticeable outliers

extending above 50 NT$. Right panel shows residual distribution with slight left skew. Bottom panel shows residuals versus log distance with a clear U-shaped pattern. This systematic pattern in residuals versus distance violates the linear assumption and indicates Model 2's flexible spline will better capture the true relationship.

Appendix B4: Convergence Trace Plots for M2 and M3

Models 2 and 3 show trace plots similar in quality to Model 1 (Figure 7), with all chains mixing well and no evidence of stuck points, divergence, or autocorrelation. All chains converge quickly to their posterior distributions, confirming that posterior inference for all three models is reliable.

Appendix B5: Spline Decomposition for Model 2

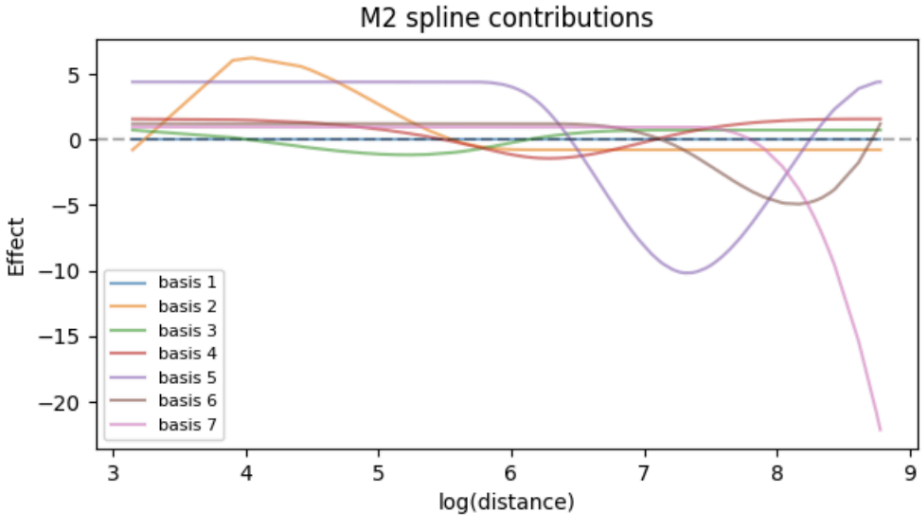


Figure B5.1. *Model 2's distance effect decomposed into seven basis function contributions (six splines plus one linear term). Each colored curve shows how much that basis function contributes to the overall distance effect. The combined effect (sum of all curves weighted by posterior coefficients) produces the smooth nonlinear distance relationship captured by the model. Notice how the contributions vary across the distance range, with some basis functions dominant near stations (left side) and others dominant far from stations (right side).*

Appendix C

Appendix C1: Cross-Validation and Pareto k Details

Leave-one-out cross-validation estimates out-of-sample predictive performance by systematically holding out each observation, fitting the model to the remaining data, and evaluating prediction accuracy. The expected log predictive density (ELPD LOO) sums these individual predictions across all observations. Higher ELPD indicates better average performance.

Table C1.1. Model Comparison Summary

Model	**ELPD LOO**	**SE**	**Rank**	**Δ ELPD**	**Δ SE**	**Problem Obs**
M2	-1348.4	23.1	1	0.0	0.0	0
M1	-1461.5	47.2	2	-113.1	45.1	1
M3	-1461.7	47.5	3	-113.3	45.1	1

The Pareto k diagnostic indicates reliability of leave-one-out estimates. Values above 0.7 suggest high leverage. Model 2's maximum k of approximately 0.2 indicates all observations have reasonable influence. Models 1 and 3 each have one observation with $k > 0.7$, but this single problematic point does not change the overall ranking given the massive ELPD gap.

Appendix C2: Detailed Effect Estimates from Model 2

Table C2.1. Model 2 Coefficient Effects

Effect	**Mean**	**89% HDI Lower**	**89% HDI Upper**	**Interpretation**
Distance: 500m vs 1000m	6.9 NT$	6.0	7.8	Per 500m closer

Age: per year older	-0.23 NT$	-0.28	-0.17	Depreciation
New vs 40-year-old	-10.2 NT$	-11.8	-8.5	Total depreciation

Appendix C3: Spatial Analysis and Geographic Effects

The heatmap in Figure 13 shows the predicted spatial interaction surface. The coefficient of 1.69 NT$ is small relative to distance effects. The spatial pattern, while visible, does not substantially improve model predictions. This suggests that the strong clustering visible in Figure 3 (Data section) is driven primarily by distance to MRT and coordinate-level variation, not by interactive geographic effects.

For practitioners seeking geographic pricing guidance, the simpler additive model (Model 1 or Model 2 structure) with latitude and longitude terms separately is preferable to the interaction approach. See Appendix C2 for coordinate coefficient estimates.

Appendix C4: Model 1 Reference Coefficients (in Real Units)

For reference, Model 1's coefficients in interpretable units (from the printed output earlier):

- Effect of 1 additional nearby store: 0.37 NT$/ping (89% HDI: 0.06, 0.68)
- Effect of 1 year older: -0.23 NT$/ping (89% HDI: -0.28, -0.17)
- Effect of moving from 1000m to 500m: 4.63 NT$/ping (89% HDI: 3.97, 5.28)

Note that Model 1's distance effect (4.63 NT$) is smaller than Model 2's (6.9 NT$), reflecting the linear model's failure to capture the steeper decline at short distances.

Appendix D

Appendix D1: Model Assumptions

All predictors were standardized before modeling to improve convergence. This means coefficients are in standard deviation units, but converting back to original units requires dividing by the original standard deviation, which was done in Section 4 for interpretation. The priors were weakly informative, centered on domain knowledge but allowing data to dominate. See Appendix B2 for prior specifications.