A Bayesian Approach to Quantitative Comparative Real Estate Valuation

Geoffrey Zhu
zyzhu@gatech.edu

Abstract

We propose a novel Bayesian hierarchical model to valuate real estate properties in Taiwan's big cities using the same data set as in [6]. By efficiently utilizing the structural information embedded in the location, the model achieves a predictive accuracy slightly surpassing the original paper while remaining highly interpretable despite the data set's small size and limited features.

1. Introduction

There is growing interest in using quantitative models to valuate real estate properties, eliminating human bias and reducing appraisal costs. However, black-box machine learning models tend to have low interpretability and demand large, comprehensive data sets to get good results.

Recently, Yeh and Hsu [6] proposed a two-stage model that seems to overcome both challenges. It operates on small datasets (e.g. 414 examples for Sindian district) derived from public records that contain limited features describing only the surroundings of the properties but lacking information on the properties themselves, such as the number of bedrooms. Nevertheless, their model is highly interpretable and can predict the unit price reasonably accurately, outperforming both linear regressions and neural networks.

We theorize that its good performance stems from the built-in inductive bias that captures the domain knowledge. The model mimics human appraisers by first estimating the impact of each feature on every property using all the data and then taking a weighted average to assign more weights to properties that look similar to the target.

It can be viewed as a form of domain-knowledge-based regularization. The first step pools all the data points to estimate the sensitivities, efficiently using the small data set. The second step takes advantage of the locality of the real estate market, avoiding underfitting. It is a manual completing pooling/partial pooling scheme that takes advantage of the structure of the data.

We argue that Bayesian models are naturally suitable for this task as it provides a principled way to combine prior knowledge and observed information from complex, structured data. It also automatically learns and applies regularization and partial pooling without ad-hoc approaches. Moreover, it produces credible intervals on the predictions and parameters, giving the users a sense of model uncertainty that is useful to make decisions.

In this paper, we develop and compare three Bayesian models and show that by efficiently utilizing the information embedded in the structures of the data, our hierarchical model surpasses the original paper's accuracy while remaining interpretable.

2. The Data

We use the same data set as was used in [6], which can be found in the UCI data collection([1]). It contains 414 examples of historical real estate valuations of the SinDian District, New Taipei City. There are six features in addition to the corresponding unit price as shown in Table 1:

	Date	Age	Distance to MRT	Number of conv stores	Lat	Lon	Unit Price
count	414.0	414.0	414.0	414.0	414.0	414.0	414.0
mean	2013.1	17.7	1083.9	4.1	25.0	121.5	38.0
std	0.3	11.4	1262.1	2.9	0.0	0.0	13.6
min	2012.7	0.0	23.4	0.0	24.9	121.5	7.6
25%	2012.9	9.0	289.3	1.0	25.0	121.5	27.7
50%	2013.2	16.1	492.2	4.0	25.0	121.5	38.5
75%	2013.4	28.1	1454.3	6.0	25.0	121.5	46.6
max	2013.6	43.8	6488.0	10.0	25.0	121.6	117.5

Table 1: Summary of data set that contains 414 examples of the historical property valuations of SinDian Distr of New Taipei.

2.1. A Qualitative Analysis of the Features

Before we start, let us analyze the features.

1. *Distance to the nearest Metro Station* Metro is a major form of transportation in the area. We theorize that the distance to the nearest MRT is essential to convenient living and hence has an inverse relationship to the value. The impact seems to possess diminishing marginal return as an additional 10-minute walk to the station feels a lot worse when the total distance is a 30-minutes than a 2-hour walk.

2. *Number of Nearby Convenience Stores*

   Having nearby convenience stores increases life quality. There is also likely some diminishing marginal utility as when there are already ten stores nearby, an additional store will not add much value.

3. *Age of House*

   In this dataset, age is the sole feature that describes the intrinsic characteristics of the properties. By common sense, newer houses are better. The impact feels not entirely linear.

4. *Transaction Date*

   The real estate market appreciated a lot during the period of study. Therefore, the transaction date is a proxy for the impact of the overall market.

5. *Latitude and Longitude of the properties*

   Our initial models consider that the other four features reflect the information embedded in location entirely; hence there is no need to include location in those models. However, location encodes other information about the neighborhood, such as social-economical status. We utilize this information in our last model.

3. Approach

We develop our models iteratively, starting from the simplest model and making revisions by identifying and addressing the weaknesses of the previous incarnations.

All the models are fitted with `pystan`. We collect 4,000 samples for inference after running 4,000 samples to warm up.

Posterior predictive checks are used to identify model deficiencies. It involves drawing samples from a trained model's posterior predictive distribution and comparing them with the observed data. Since they represent the replicate data that we could have observed, they should have statistical properties similar to our observations ([2]) if the model is right. Posterior predictive checks is a powerful way to reveal model weaknesses.

Out-of-sample predictive performance is evaluated both through approximated cross-validation and external validation. Following the original paper, we divide the dataset randomly into the training set (2/3) and the test set (1/3). During development, we use PSIS [1] and WAIC [2] to gauge out-of-sample predictive performance. Both are approximations of leave-one-out CV on the logarithmic score ([4, 2]). Watanabe ([5]) recommends computing and contrasting both to monitor unreliability. In the end, we measure predictive performance on the test set on metrics specific to the real estate valuation and compare them to the same metrics in the original paper.

[1] Pareto-smoothed Importance Sampling Cross Validation

[2] Widely Applicable Information Criterion

4. Experiments

4.1. Bayesian Linear Regression

4.1.1 The Model

Equation 2 - 7 below shows the linear regression model with somewhat informative priors based on our analysis in Section 2.1.

$$y_i \sim N(\mu_i, \sigma^2), \text{ where } i = 1, \ldots, n \tag{1}$$
$$\sigma \sim Exponential(1) \tag{2}$$
$$\mu_i = \alpha + \sum_{k=1}^{4} \beta_k x_{ik} \tag{3}$$
$$\ln(-\beta_1) \sim N(0, 1.33) \tag{4}$$
$$\ln(-\beta_2) \sim N(0, 1.33) \tag{5}$$
$$\ln(\beta_3) \sim N(0, 1.33) \tag{6}$$
$$\beta_4 \sim N(0, 20) \tag{7}$$

, where β_k, where $k = 1, \ldots, 4$, is coefficients for 1) logarithmic of distance to Metro, 2) logarithmic of age, 3) logarithmic of number of nearby convenience stores, and 4) transaction date respectively. We standardize the features to have mean 0 and standard deviation 1 on the training set.

As discussed in Section 2.1, we can determine that $\beta_1, \beta_2 < 0$ and $\beta_3 > 0$ before we see the data and therefore we use lognormal distributions with reasonable standard deviations as their priors. For transaction date, since we know little about the market price during the period, we use a vague normal distribution as the prior.

4.1.2 Fitting the Model

We use `pystan` to perform MCMC on the model and collect 4,000 samples for inference after warming up with 4,000 samples.

Table 2: Parameters for the Linear Regression

	mean	sd	hdi_3%	hdi_97%
β_1 (log(dst MRT))	-9.136	0.738	-10.554	-7.789
β_2 (log(age))	-2.664	0.553	-3.748	-1.664
β_3 (log(# stores))	1.531	0.713	0.207	2.783
β_4 (date)	2.478	0.559	1.446	3.557
α (intercept)	38.122	0.516	37.140	39.056
σ 8.799	0.376	8.104	9.499	

Table 2 summarizes the posterior distributions of the parameters. As zero is outside the credible intervals of all four

coefficients, we conclude that all four significantly impact the price. The coefficients for log(distance to Metro) and log(house age) are highly negative. In contrast, that of log(# stores) is highly positive, indicating the closer to the Metro station, the newer the property, and the more nearby convenience stores, the higher the unit price. These all agree with our prior analysis in Section 2.1. Finally, the coefficient for the transaction date is also positive, indicating a rising housing market.

Table 3 shows the approximate out-of-sample predictive performance using PSIS and WAIC. Both metrics estimate the leave-one-out cross-validation performance on the log score and can be used to compare with other models.

Table 3: Predictive Performance of the Linear Regression

	value	se	p
PSIS	-1007.27	40.09	14.53
WAIC	-1007.33	40.16	14.60

4.1.3 Posterior Predictive Check

Bayesian models are generative. By drawing samples from the posterior predictive distribution of a fitted model, we can generate replicate data from the model, i.e. data that could have been observed given the model. One important check is therefore to compare the replicate data with the original data.

In Figure 1, we draw multiple sets of the replicate data from the posterior predictive distribution and plot the distribution of each set as a red line, overlaying on the distribution of the actual observations. The mean of all the red lines is plotted as the dotted blue line.

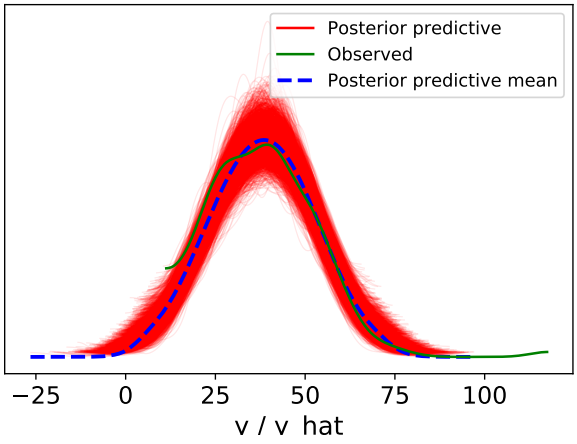

Figure 1: Posterior predictive check. The figure shows the distribution of replicate data points drawn from the posterior predictive distribution, overlayed on the observed data.

The vertical thickness of the red line cloud reflects the uncertainty of the prediction. It is highest around the peak (when y is around 40). The model is more confident about its predictions at 20 and 60. It is good news that the observed data mostly falls inside the predicted range of red lines because it indicates the predictions are plausible.

However, on the right wing, we see the model suffers over-dispersion. The actual observed data has a fatter tail than the model prediction. As a result, the data points on the right tail have too strong an influence on the model. We can see that the whole mean prediction (dotted line) is pulled to the right.

One possible reason for the over-dispersion is that the features cannot account for all the variations of prices. Nothing in the feature distinguishes between a luxury 3-bedroom condo and a shabby studio. In regions where properties are mixed, we can imagine that unit prices are generated by different tiers of the properties, causing a fat tail.

4.2. Robust Bayesian Linear Regression

To handle the over-dispersion that we identified in the posterior predictive check, we formulate a robust regression by replacing the normal distribution in the likelihood in Equation 2 with a Student's t-distribution, which is known to have fatter tails. Equation 9 shows the new likelihood:

$$y_i \sim t(\mu_i, \sigma^2, \nu), \text{ where } i = 1, \ldots, n \tag{8}$$

$$\nu \sim Gamma(2, 0.1) \tag{9}$$

Here ν is the degree of freedom for the Student's t-distribution. The smaller the value, the fatter the tails. When $\nu \to +\infty$, the Student-t distribution converges to a normal distribution. This is to say that the previous model is nested in the new model. Following [3], we give ν a vague prior $Gamma(2, 0.1)$ so that it is free to take most values but mostly stays above 2, which is needed for the mean to exist.

4.2.1 Fitting the Model

Table 4: Parameters of the Robuster Linear Regression

	mean	sd	hdi_3%	hdi_97%
β_1 (log(dst MRT))	-8.655	0.595	-9.750	-7.541
β_2 (log(age))	-2.811	0.514	-3.771	-1.873
β_3 (log(# stores))	2.152	0.656	0.901	3.343
β_4 (date)	1.895	0.428	1.122	2.708
α (intercept)	37.081	0.416	36.312	37.865
σ	5.614	0.445	4.814	6.493
ν (degree of freedom)	3.660	0.934	2.197	5.355

Table 4 summarizes the posterior distributions of the robust regression model. Compared to Table 2, we see that the

β's have slightly shifted. What is most worth noting is the new parameter ν, the degree of freedom for the Student's t-distribution. The mean value is 3.660, showing that the model now captures the very fat tail!

Table 5: Predictive Performance of the Robust Linear Regression

	value	se	p
PSIS	-964.62	18.06	9.10
WAIC	-964.48	17.99	8.96

Table 5 shows the PSIS and WAIC. Compared with the previous model (Table 3), the out-of-sample predictive power is better as both PSIS and WAIC have improved by about 5%.

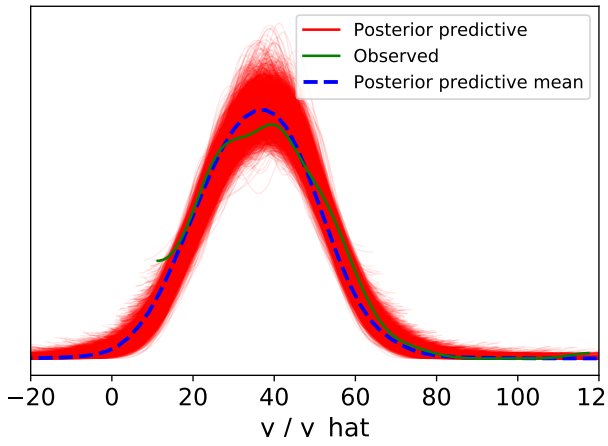

Figure 2: Posterior predictive check for robust regressions

A look at the posterior prediction check is even more revealing. In Figure 2, the right tails of the red lines are indeed much fatter and mostly covers the observed data. Also, the predicted values (dotted line) have shifted left a bit as the data points on the far right side now have less influence.

4.3. Bayesian Hierarchical Model

With only six features, how do we further improve our model? One way is to utilize the information embedded in the internal structures fully.

Location, location, location!

Real estate truism says "location, location, location!" When we previously modeled the properties, we have assumed the other four features have accounted for all information from location. In reality, other things, such as local density, culture, and social-economic status, can be tied to the neighborhood and may also affect property price. Different neighborhoods may even value things such as the number of nearby convenience stores differently.

To model this, we use *k-means* to divide the area into five regions (clusters), which can then have their individual intercepts and regression coefficients to reflect the location difference. These regional parameters can influence one another through shared global-level hyperpriors, achieving partial pooling. This is the hierarchical model's way of recognizing both the differences and commonalities among clusters to best use information.

Influence of the housing market is multiplicative

When the real estate market goes up 10%, the prices of all properties are likely to go up by 10%. This is to say that the effect of the transaction date is best modeled multiplicatively. Let r be the annualized average continuous compounding return for the period, we can model the effect of the date as (base price) $\times\, e^{r\Delta t}$, where Δt is the time of transaction since the start of the period.

To put these two improvements together, we formulate a Hierarchical Model as follows:

$$y_i \sim t(\mu_i, \sigma^2, \nu), \text{ where } i = 1, \ldots, n \tag{10}$$
$$\nu \sim Gamma(2, 0.1) \tag{11}$$
$$\sigma \sim Exponential(1) \tag{12}$$
$$\mu_i = \left(\alpha_{\text{CLUSTER}[i]} + \sum_{k=1}^{3} \beta^T_{\text{CLUSTER}[i]} x_{ik}\right) \times e^{r(t_i - t_0)} \tag{13}$$
$$r \sim N(0, 0.2) \tag{14}$$
$$\beta_{jk} \sim N(\beta_k, \sigma_k) \tag{15}$$
$$\ln(-\beta_1) \sim N(\beta_1, 1.33) \tag{16}$$
$$\ln(-\beta_2) \sim N(0, 1.33) \tag{17}$$
$$\ln(\beta_3) \sim N(0, 1.33) \tag{18}$$
$$\sigma_k \sim Exponential(1) \tag{19}$$

Here, CLUSTER$[i] \in \{1, \ldots, J\}$, where $J = 5$, gives the cluster membership of example i, r is the average annualized rate of return for the time period, and t_0 is the start of the time period. The parameters for cluster j are α_j and β_{jk} respectively and α and β_k are global-level hyperparameters.

Two things are worth noting: 1) now that we model the effect of transaction date using the rate of return, it is easier to specify a meaningful prior. In Eq 14, we set $r \sim N(0, 0.2)$ as from common sense we know real-estate market price usually change less than 20% a year. It makes the prior sensibly vague; 2) in Eq 15, the σ_k's control how tightly the cluster-level parameters are bound to the global level hyperparameters. It controls the degree of partial pooling and is the key to adaptive regularization. We learn these parameters from the data.

Table 6 shows the parameters of the fit model. We have grouped the coefficients by feature.

As we can see, both the intercept and coefficients vary quite a lot among clusters. As previously discussed,

Table 6: Parameters of the Hierarchical Model

	mean	sd	hdi_3%	hdi_97%
β_1	-5.915	1.234	-8.273	-3.606
β_{11}	-4.986	1.417	-7.299	-2.212
β_{12}	-6.359	1.652	-9.836	-3.401
β_{13}	-6.100	0.991	-7.833	-4.097
β_{14}	-6.676	0.999	-8.632	-4.865
β_{15}	-6.953	1.643	-10.301	-4.199
σ_1	1.286	0.982	0.060	3.064
β_2	-2.360	1.074	-4.016	-0.173
β_{21}	-2.352	0.696	-3.687	-1.075
β_{22}	-2.138	1.036	-4.147	-0.302
β_{23}	-6.068	0.557	-7.060	-4.961
β_{24}	-2.079	0.623	-3.206	-0.957
β_{25}	-1.985	1.932	-5.334	1.920
σ_2	2.122	0.806	0.920	3.653
β_3	1.280	0.757	0.098	2.670
β_{31}	0.615	0.770	-0.710	2.173
β_{32}	1.298	1.673	-1.785	4.705
β_{33}	0.237	1.227	-2.125	2.421
β_{34}	3.013	1.001	1.127	4.857
β_{35}	2.128	1.343	-0.303	4.763
σ_3	1.544	0.819	0.113	3.017
α	34.335	1.897	30.687	37.730
α_1	37.625	1.019	35.722	39.651
α_2	30.749	1.873	27.502	34.633
α_3	39.117	1.190	36.851	41.401
α_4	31.750	0.851	30.264	33.488
α_5	32.029	2.994	26.162	37.645
σ_α	3.525	1.076	1.805	5.575
σ	4.495	0.350	3.869	5.165
ν	3.453	0.790	2.131	5.050
r (annual return)	0.177	0.033	0.117	0.242

$\sigma_1, \sigma_2, \sigma_3, \sigma_\alpha$ control the strength of partial pooling. Different features have different values, indicating that adaptive regularization is working. We also note that the intercept α is no longer aligned to the sample mean of the target price. It is because we have now taken out the effect of market appreciation. The model has estimated that the market appreciated $r = 17.7\%$ for the year. Modeling the effect of transaction date this way for sure increases interpretability.

Table 7 shows the estimated out-of-sample predictive performance. We see another 6% improvement over the last model. It is worth noting that although we added 30 real parameters, we see the effective parameters increased only about 10, indicating a pretty strong shrinkage toward the global hyperparameters.

Table 7: Predictive Performance of the Robust Linear Regression

	value	se	p
PSIS	-912.32	18.55	19.80
WAIC	-912.24	18.53	19.71

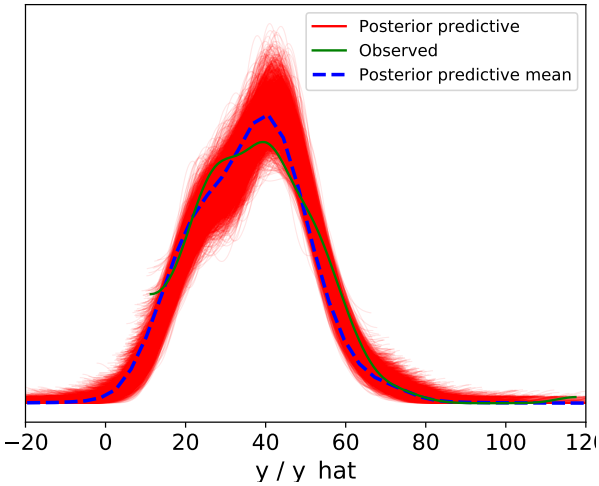

Figure 3: Posterior predictive check for the Hierarchical Model

Figure 3 shows the posterior predictive check. The replicated data shows increased agreement with observed data. There are also more details in the prediction as the predictions are no longer symmetrical, smooth curves. It is because the model now considers location differences. Finally, the model seems to have a better sense of prediction uncertainty.

4.4. Interpretability

Our models have great interpretability. First, all the inputs and parameters have physical meanings that the users can understand and do what-ifs. For example, it is intuitive to show that since the market has appreciated 17.7% for the year, we have to adjust up old valuations from months ago in a prorated fashion. Second, the hierarchical model is likely how users think of real estate properties and may help the users understand why a specific property has a certain price. Users have innate insights about the differences in neighborhoods, and may appreciate how the model assigns different values for each feature by region. For example, it is helpful to show the users that the model thinks neighborhood A values convenience stores more than neighborhood B. Finally, the model assigns credible intervals to the parameters and predictions. By letting the users know where the model is less confident, it adds transparency and helps the users to

make decisions.

4.5. Performance on Test Set

Table 8 shows the predictive performance of all three models on the test set. We have also included the published numbers in the original paper for their models. We use RMSE and 10%/20% hit rate as the metric following the original paper. Compared to the log score that we have been using for approximate cross-validation, these metrics represent the things the users care about and thus are important to examine.

The out-of-sample performance of our three models follows the same trend as predicted by PSIS and WAIC. All the models have good performance. The Hierarchical Bayesian model appears to perform the best among all the models.

Table 8: Comparison of Test Set Performance of All Models

Model	RMSE	20% error hit rate	10% error hit rate
Linear regression	7.65	0.73	0.44
Robust regression	7.39	0.79	0.45
Hierarchical	**7.26**	**0.82**	**0.56**
Custom Model (original paper)	7.73	0.802	0.529
Neural Net (original paper)	8.06	0.78	0.488

5. Future Work

There are many other opportunities to improve the model further. One possibility is to incorporate better domain knowledge. For instance, we imagine that, instead of dividing up the area using *k-means*, people with intimate knowledge about the area can divide the regions much more intelligently.

Another possibility is to continue loosening simplifying assumptions to fit the reality better. For instance, we have modeled the market rate of return as a constant throughout the period. To allow for market price fluctuations, we can model it as 12 monthly returns, which are not correlated according to the efficient market theory and hence can be modeled as independent Gaussian distributions with a common standard deviation.

6. Conclusion

In this paper, we developed and compared three Bayesian real estate valuation models using the same data set as used by [6]. Our best model surpassed the original paper in predictive accuracy while maintaining good interpretability.

The Bayesian method provides a principled way to incorporate prior knowledge from the priors and the structures of the model, and a flexible framework to model data that contain diverse structures. It efficiently uses information from structured data and adaptively regularizes the trade-off between underfit and overfit. In our last model, such flexibility allows us to capture the information embedded in the location clusters to achieve good performance despite the small data set and incomplete features. Additionally, Bayesian models naturally account for the model uncertainty, providing the users with credible intervals to help them make informed decisions. We believe our Bayesian model is a competitive approach to valuating the real estate market in big cities in Taiwan.

References

[1] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[2] Andrew Gelman et al. *Bayesian Data Analysis*. Zeroth. Chapman and Hall/CRC, Nov. 2013. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

[3] Miguel A. Juárez and Mark F. J. Steel. "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions". In: *Journal of Business & Economic Statistics* 28.1 (Jan. 2010), pp. 52–66. ISSN: 0735-0015. DOI: 10/cj94c7.

[4] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. First. Chapman and Hall/CRC, Jan. 2018. ISBN: 978-1-315-37249-5. DOI: 10.1201/9781315372495.

[5] Sumio Watanabe. *Comparison of PSIS Cross Validation with WAIC*. URL: http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/psiscv.html.

[6] I-Cheng Yeh and Tzu-Kuang Hsu. "Building Real Estate Valuation Models with Comparative Approach through Case-Based Reasoning". In: *Applied Soft Computing* 65 (Apr. 2018), pp. 260–271. ISSN: 15684946. DOI: 10/gnkwjz.