



RESEARCH ARTICLE

WILEY

The role of investor sentiment in forecasting housing returns in China: A machine learning approach

Oguzhan Cepni^{1,2}  | Rangan Gupta³  | Yigit Onay²

¹Copenhagen Business School, Frederiksberg, Denmark

²Central Bank of the Republic of Turkey, Ankara, Turkey

³Department of Economics, University of Pretoria, Pretoria, South Africa

Correspondence

Oguzhan Cepni, Copenhagen Business School, Frederiksberg DK-2000, Denmark.
Email: oce.eco@cbs.dk

Abstract

This paper analyzes the predictive ability of aggregate and disaggregate proxies of investor sentiment, over and above standard macroeconomic predictors, in forecasting housing returns in China, using an array of machine learning models. We find that our new aligned investor sentiment index has greater predictive power for housing returns than the principal component analysis (PCA)-based sentiment index, used earlier in the literature. Moreover, shrinkage models utilizing the disaggregate sentiment proxies do not result in forecast improvement indicating that aligned sentiment index optimally exploits information in the disaggregate proxies of investor sentiment. Furthermore, when we let the machine learning models to choose from all key control variables and the aligned sentiment index, the forecasting accuracy is improved at all forecasting horizons, rather than just the short-run as witnessed under standard predictive regressions. This result suggests that machine learning methods are flexible enough to capture both structural change and time-varying information in a set of predictors simultaneously to forecast housing returns of China in a precise manner. Given the role of the real estate market in China's economic growth, our result of accurate forecasting of housing returns has important implications for both investors and policymakers.

KEYWORDS

Bayesian shrinkage, housing prices, investor sentiment, time-varying parameter model

1 | INTRODUCTION

Financialization of the housing market, that is, its treatment as a commodity, though a contentious issue due to housing's primary role to serve as a social good, is now a well-established global fact (Aalbers, 2016). And in this regard, China, a major player in the world economic system, is not far behind, though the process started more recently compared with the western world (Wu et al.,

2020). The commodification of the housing market is considered to be one of the major drivers of China's economic development (Hsing, 2010; Lin, 2014; Tao et al., 2010; Wu, 2015) and is believed to have played an important role in crisis management. For instance, the suspension of welfare housing provision in 1998 as a response to the 1997 Asian financial crisis (Logan et al., 2010), and more recently, the housing boom triggered by the injection of 4 trillion yuan into infrastructure and urban

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of Forecasting* published by John Wiley & Sons Ltd.

development after the 2008 global financial crisis (Deng & Chen, 2019).¹ Put alternatively, financialization of the housing market in China (and around the world) basically implies that price comovement between real and financial assets, that is, the equity market (as well as between different assets), is likely to increase (Hong & Li, 2019a), with the underlying driver being investor sentiment (Tang & Xiong, 2012). Moreover, while documenting the relatively large amount of existing evidence of investor sentiment in affecting Chinese stock returns, Su et al. (2020) point out that, because individual investors account for a large proportion of Chinese stock investors, they not only tend to make irrational trades in the stock market but are also more likely to carry their emotions to other markets.

Now, the importance of the real estate markets for China's past and continued economic growth and the key overall role of house prices as a leading indicator of the macroeconomy is well recognized (Chow et al., 2018).² This is especially due to the introduction of neoliberal reforms in the 1990s, and particularly since 1998, when public sector housing allocation was replaced by market allocation and quasi-privatization of property (Theurillat et al., 2016). Hence, accurate forecasting of housing returns based on the information content of investor sentiment in the wake of financialization is an important question for policy authorities to gauge the future path of the overall domestic economy. Given China's position in the global economy as the second-largest economy (after the United States), with its share of global gross domestic product (GDP) adjusted for purchasing-power-parity (PPP) being 19.72% (Schwab, 2019), performance of the Chinese economy is also a pertinent issue for policymakers around the world.

Against this backdrop, the objective of our paper is to forecast composite housing returns for 70 large and medium-sized cities in China over the monthly period of 2011:01 to 2018:12, given an in-sample period of 2006:01 to 2010:12, based on investor sentiment (controlling for

other predictors) using a variety of machine learning methods (such as generalized approximate message passing [GAMP], Bayesian model averaging, Ridge regression, least absolute shrinkage operator [LASSO], and elastic net [ENET]). These shrinkage-based approaches allow us to efficiently conduct the forecasting experiment, without suffering from the “curse of dimensionality,” especially in the context of a time-varying framework with multiple predictors and relative short span (13 years) of data (in our case 156 monthly observations).³ Our paper can be considered to be an extension of Hong and Li (2019b), whereby they use wavelet analysis to provide in-sample evidence of the predictability of housing returns in China due to investor sentiment. However, because in-sample predictability does not guarantee forecasting gains, and as pointed out by Bork and Møller (2015) that the ultimate test of any predictive model (in terms of the econometric methodologies and the predictors used) is in its out-of-sample performance, evidence of forecastability, if it exists, would provide more robust evidence (relative to an in-sample analysis) of the role of investor sentiment for future housing returns.⁴

One must recall that investor sentiment is a latent variable and needs to be derived from appropriate proxies. Given this Hong and Li (2019b), followed Baker and Wurgler (2006) to form the investor sentiment index using the principal component analysis (PCA) to aggregate the information from six individual proxies (the closed-end fund discount; average first-day returns on initial public offerings [IPOs]; the ratio of the number of advancing stocks to the number of declining stocks; new A-share market accounts; market turnover rate; and CCI), which we use as well, both as an index and individually. But as an alternative to PCA, we also use partial least squares (PLS) to construct the sentiment index. Econometrically speaking, the first principal component is indeed the best combination of the six proxies that represents the highest percentage of the total variations of the proxies. Because all the proxies may have approximation errors to the true but unobservable investor sentiment, and these errors are parts of their variations, the first principal component can potentially contain a

¹Similarly, mortgage-backed securities and derivatives with opaque structures, excessive leverage, and poor risk management all had a part in the collapse of the US housing market in 2008 and damaged the global economy, leading born of the concept of “systemic risk,” or the possibility that a single event, like the bankruptcy of a major financial institution, may devastate financial markets and the whole economy (see, Bullard et al., 2009; Eichengreen et al., 2012; Gorton, 2009; Sanders, 2008). As a matter of fact, this important crisis event has been a key factor that has caused the popularity of searching for ways to forecast house prices and related variables (e.g., early warning systems for US house prices) (Kouwenberg & Zwinkels, 2014).

²Using a Granger causality approach, Luo et al. (2007) find that relationships between macroeconomic variables and house prices are unstable over time and location, causing difficulty in forecasting of house prices.

³In the context of the housing market, some other studies utilize machine learning techniques (e.g., Hausler et al., 2018; Park & Bae, 2015; Rodriguez Gonzalez et al., 2022).

⁴Furthermore, there are a lot of relevant empirical studies have been published focusing on the ability of sentiment indicators to predict house prices or related variables for other countries as well (see, among others, Clayton et al., 2009; Croce & Haurin, 2009; Dietzel et al., 2014; Hausler et al., 2018; Kunze et al., 2020; Marcato & Nanda, 2016; McLaren & Shanbhogue, 2011; Rodriguez Gonzalez et al., 2018; Rodriguez Gonzalez et al., 2022; Tsolacos, 2012).

substantial amount of common approximation errors that are possibly not relevant for forecasting asset returns (Bai & Ng, 2008; Boivin & Ng, 2006). Given this, we align the investor sentiment measure with the purpose of explaining the housing returns by extracting the most relevant common component from the proxies. In other words, we separate out information in the proxies that is relevant to the expected housing returns from the error or noise, by using the PLS method originally developed by Wold ((1966), (1975)), and applied to housing return and financial data more recently by Kelly and Pruitt ((2013), (2015)), Bork and Møller (2018), and Cepni et al. (2020). Hence, the usage of the PLS to obtain an aligned investor sentiment index can also be considered as a contribution of our study, besides the full-fledged out-of-sample forecasting exercise using machine learning methods.

To the best of our knowledge, this is the first attempt to forecast housing returns in China using aggregated and disaggregated proxies of investor sentiment and other macroeconomic and financial variables, based on a wide array of machine learning methods. The two papers that we could find which have produced out-of-sample forecasting of housing returns for China is that of Wei and Cao (2017) and Salisu and Gupta (2021). While the latter paper shows that monthly disaggregated oil shocks, that is, supply, global economic activity, oil-specific demand, and oil inventory demand, can be used to forecast quarterly housing returns of China based on a mixed-frequency model, the former paper highlights the role of a Google search index (associated with city name plus house price), instead of fundamental macroeconomic or monetary indicators, based on a dynamic model averaging (DMA) framework.

Our paper also contributes to the growing literature that examines the role of sentiment on asset prices, especially in stock markets (Fisher & Statman, 2000; Gao et al., 2020; Huang et al., 2015; Jiang et al., 2019). Due to the lacking of sentiment measures, empirical studies on the relationship between sentiment and housing returns are remarkably scarce, even though individual investors dominate the housing market with limited access to complete information, making them more susceptible to market sentiment. Short selling limits, high transaction costs, and more extensive information asymmetries in the housing market may all contribute to the creation of persistent arbitrage possibilities while also limiting the capacity of informed traders to minimize mispricing when it arises in the market (Glaeser et al., 2014). Hence, such limitations provide a unique environment to assess the predictability of sentiment on house prices. Given this gap, we contribute to this literature by showing that housing sentiment includes essential information to

predict the house price changes in the Chinese housing market.

The remainder of the paper is organized as follows: Section 2 presents the data; Section 3 outlines the methodologies used, with Section 4 presenting the main econometric results along with robustness analysis, and Section 5 concludes the paper.

2 | DATA

Housing price is downloaded for 70 large- and medium-sized cities in China over the monthly period of 2006:01 to 2018:12.⁵ Then, the composite housing price index is computed as the average of these indices. The monthly housing price return calculated as using the formula: $Houret_t = \ln(HPI_t) - \ln(HPI_{t-1}) \times 100$ where HPI_t is the monthly house price index at time t . We construct our sentiment index using the six individual sentiment proxies based on the work of Hong and Li (2019b)⁶. Following Baker and Wurgler (2006), they form the investor sentiment index using PCA (Investor.Sent.PCA) to aggregate the information from six individual proxies: the closed-end fund discount (Dcef), average first-day returns on IPOs (RIpo), ratio of the number of advancing stocks to the number of declining stocks (Adrt), new A-share market accounts (NA), market turnover rate (Turn), and consumer confidence index (CCI).⁷ As an alternative to the PCA, we use PLS to construct the sentiment index. Finally, we also collect a set of key economic variables which are growth of industrial production (IP), consumer price index inflation (CPI), the People's Bank of China's policy rate (IR), and returns of the Shanghai composite stock market index (SMR). Raw values of all control variables are downloaded from the Bloomberg terminal. Table 1 presents the summary statistics of our variables.

3 | METHODOLOGIES

3.1 | The construction of a new sentiment index

To construct our investor sentiment index (Investor.Sent.PLS) using the information contained in each of six individual sentiment proxies, we employ the PLS method to the same six proxies. In particular, we utilize the PLS

⁵Data can be downloaded from the official website of China's National Bureau of Statistics: <https://data.stats.gov.cn/>.

⁶The data of sentiment proxies ends on December 2018. As a result, the sample period used in our analysis ends on that date.

⁷We thank Dr. Yun Hong and Dr. Yi Li for sharing their data in this regard.

TABLE 1 Descriptive statistics

Variable	Houret	SMR	IR	IP	CPI	Dcef	RIpo	Adrt	NA	Turn	CCI
Mean	0.354	0.448	2.991	-0.078	0.000	0.000	0.044	0.050	7.512	0.021	0.036
Median	0.352	0.833	2.841	-0.094	0.000	0.009	-0.047	-4.423	-9.501	-0.021	-0.102
Std. Dev	0.475	6.906	1.157	1.352	0.580	0.038	0.698	15.313	47.960	0.138	2.309
Kurtosis	0.538	1.132	2.337	8.850	2.751	-0.038	31.234	19.800	8.935	2.081	8.053
Skewness	-0.226	0.130	1.073	-0.703	-0.670	-0.412	4.443	4.078	2.860	1.486	1.504
Range	2.807	39.660	7.255	12.658	4.091	0.197	6.668	114.963	287.711	0.743	18.922
Min.	-1.008	-19.396	0.880	-8.066	-2.608	-0.110	-0.840	-9.356	-37.722	-0.185	-5.192
Max.	1.799	20.264	8.135	4.592	1.483	0.088	5.828	105.607	249.989	0.558	13.730

method using Friedman et al.'s (2001) two-step approach. The algorithm starts by standardizing each proxies x_j ($j = 1, \dots, p$) to have a zero mean and unit variance. Then, univariate regression coefficients $\widehat{\gamma}_{1j} = \langle x_j, y \rangle$ are computed for each j . From this, the first PLS direction $z_1 = \sum_j \widehat{\gamma}_{1j} x_j$ is obtained as the weighted sum of the vector

of univariate regression coefficients and the original set of sentiment proxies. Hence, the construction of the PLS direction takes into account the degree of association between housing returns and common factors. In the following step, the “target” variable y is regressed on z_1 , resulting in a coefficient θ_1 , and then, all inputs are orthogonalized with respect to z_1 . This process is iterated until PLS produces a sequence of $l < p$ orthogonal directions.

Because PLS utilizes the housing returns to construct the directions, its solution path is a nonlinear function of housing returns. While PCA finds directions that maximize only the variance, PLS aims for the directions that have high variance and high correlation with the target variable which intuitively could increase the forecasting power of a PLS-based index compared with a PCA-based index.

More specifically, the m th PLS direction γ_m solves the following optimization problem:

$$\begin{aligned} & \max_{\alpha} \\ & \text{Corr}^2(y_t, X_{\alpha}) \text{Var}(X_{\alpha}), \\ & \text{subject to} \\ & \|\alpha\| = 1, \alpha' S \widehat{\gamma}_l = 0, l = 1, \dots, m-1, \end{aligned} \quad (1)$$

where S represents the sample covariance matrix of the x_j . We choose the first common component as a new investor sentiment index, which efficiently incorporates all the relevant information from the each of the six sentiment proxies for housing returns.

3.2 | Time-varying parameter regressions: Machine learning approaches

After the construction of new sentiment index, this section introduces a comprehensive list of competing specifications and estimation algorithms which are presented in the following subsections. All models are estimated on an expanding window using only information available at the time of forecast. In addition to standard shrinkage methods such as ridge regression, LASSO, and ENET, we implement Bayesian model averaging methods and a recently developed algorithm of GAMP. The main advantage of this algorithm is that unlike the existing posterior simulation techniques, which are unable to scale up to large dimensions because of the computational inefficiency and increased numerical inaccuracy associated with repeated sampling using Monte Carlo methods of sampling, the GAMP algorithm provides a “faster” posterior inference and can be used in high-dimensional setting (Korobilis, 2021).

3.2.1 | GAMP algorithm

Although the Bayesian approach using Markov chain Monte Carlo (MCMC) methods is a powerful tool to take into account the changing nature of the relationship between variables, computational concerns with these methods, as well as large errors related to repeated sampling through Monte Carlo, make it harder to rely on them in case the dimension of the econometric model is high (Angelino et al., 2016). As an alternative to computational limitations, message passing algorithms come to the forefront representing a highly efficient and easy-to-implement Bayesian estimation algorithm which makes it possible to take stochastic volatility and parameter instability into account with a large set of predictors. Moreover, unlike “well-established” MCMC

algorithms, GAMP-based algorithms require minimal or no tuning.

TVP-GAMP offers an efficient way of estimation by allowing time-varying parameter and variable selection together with stochastic volatility simultaneously with no restrictions on the number of predictors. TVP-GAMP procedure relies on the visualization of the relation between random variables within the framework of factor graphs.⁸ By factorizing joint posterior distribution functions of random variables into smaller parts, the procedure offers a localized way to iteratively approximate complex set of conditional marginal distributions with an approach called the sum-product algorithm, which is also known as ‘‘Belief Propagation.’’⁹

To illustrate the estimation process through GAMP procedure, a time-varying parameter specification with stochastic volatility is given as follows:

$$y_t = x_t \beta_t + \varepsilon_t, \quad (2)$$

subject to an initial condition for β_t at $t=0$, where y_t is the t th observation on the dependent variable, $t=1, \dots, T$; x_t is a $1 \times q$ vector of predictors, β_t is a $q \times 1$ vector of coefficients, and $\varepsilon_t \sim N(0, \sigma_t^2)$ with σ_t^2 the time-varying variance parameter.

Equivalently, rewriting the regression in the static form, we have the matrix representation as follows:

$$y = \mathcal{X}\beta + \varepsilon, \quad (3)$$

where $y = [y_1, \dots, y_T]'$ and $\varepsilon = [\varepsilon_1, \dots, \varepsilon_T]'$ are column vectors representing the observations y_t and ε_t , respectively, $\beta = [\beta_0', \beta_1', \dots, \beta_T']'$ is a $(T+1)q \times 1$ vector containing the coefficients. The number of parameters to be estimated for the coefficients is equal to $q = (T+1)p$ which is not possible to estimate with the classical OLS procedure. State space models have been offered in the literature to overcome the problem of identification such that a stochastic process, most typically random walk (RW), is assumed for the coefficients and the estimation is accomplished through MCMC methods for these models.

As an alternative, TVP-GAMP attempts to identify the whole set of coefficients in the TVP regression with data-driven hierarchical shrinkage priors. Consider i.i.d prior $p(\beta) = \prod_{i=1}^q p(\beta_i)$,¹⁰ then the marginal posterior for

$\beta_i, i=1, \dots, q$ obtained through Bayes theorem requires evaluation of a $(q-1)$ -dimensional integral of the form

$$\begin{aligned} p(\beta_i|y) &= \int p(\beta|y) d\beta_{j \neq i} \\ &= \int p(y|\beta) p(\beta) d\beta_{j \neq i} \\ &= p(\beta_i) \int p(y|\beta) \prod_{j=1, j \neq i}^q p(\beta_j) d\beta_{j \neq i}. \end{aligned} \quad (4)$$

Computation of the above summation can be quite cumbersome especially in case of high dimensionality problem in parameters. Factorizing the above-mentioned posterior distribution through factor graphs, we define $\mu_{p(\cdot) \rightarrow a}$ the message passed from probability function $p(\cdot)$ to random variable a , then

$$p(\beta_i|y) = \mu_{p(\beta_i) \rightarrow \beta_i} \prod_{t=1}^T \mu_{p(y_t|\beta) \rightarrow \beta_i}, \quad (5)$$

where $\mu_{p(\beta_i) \rightarrow \beta_i} = p(\beta_i)$. According to sum-product rule, we further have

$$\mu_{p(y_t|\beta) \rightarrow \beta_i} = \int p(y_t|\beta) \prod_{j=1, j \neq i}^p \mu_{\beta_j \rightarrow p(y_t|\beta)} d\beta_{j \neq i}, \quad (6)$$

$$\mu_{\beta_j \rightarrow p(y_t|\beta)} = p(\beta_j) \prod_{s=1, s \neq t}^T \mu_{p(y_s|\beta) \rightarrow \beta_j}. \quad (7)$$

We can estimate Messages 6 and 7 above using the following iterative scheme, for $r=1, \dots, R$ where r denotes the order of iteration.

$$\mu_{p(y_t|\beta) \rightarrow \beta_i}^{(r+1)} = \int p(y_t|\beta) \prod_{j=1, j \neq i}^p \mu_{\beta_j \rightarrow p(y_t|\beta)}^{(r)} d\beta_{j \neq i}, \quad (8)$$

$$\mu_{\beta_j \rightarrow p(y_t|\beta)}^{(r+1)} = p(\beta_j) \prod_{s=1, s \neq t}^T \mu_{p(y_s|\beta) \rightarrow \beta_j}^{(r)} \quad (9)$$

The GAMP algorithm employs Gaussian and Taylor series approximations which are based on asymptotic results such that as the number of parameters to be estimated increases, the analytical solutions derived from above iterations for the first two moments of the predictors become more reliable.¹¹ As in the case of the coefficient estimation of the exogenous predictors, stochastic volatility (σ_t^2) estimation is accomplished through data-

⁸Factor graph approach decomposes random variables into quantities of lower dimensions. See Korobilis (2021) for simplified illustration of the factor graph representation.

⁹See, for example, Pearl (1982) for details.

¹⁰Sparse Bayesian learning (SBL) prior as described in Tipping (2001) as hierarchical priors is used, because it has desirable variable selection properties (Korobilis, 2013).

¹¹For an illustration of a simplified version of GAMP algorithm to derive mean and variance of β , see Korobilis (2021).

driven priors without resorting to any form of Markov-based dependence to σ_{t-1}^2 .

3.2.2 | Bayesian model averaging (TVP-BMA)

We employ time-varying parameter Bayesian model averaging approach of Groen et al. (2013) which incorporates model uncertainty as the relationship between housing returns and predictor variables is likely to have changed over time (Wei & Cao, 2017).

In particular, the TVP-BMA specification takes the following form:

$$\begin{aligned} y_{t+h} &= \sum_{j=1}^p x_{jt} s_j \beta_{jt} + \varepsilon_{t+h}, \\ \beta_t &= \beta_{t-1} + \eta_t, \end{aligned} \quad (10)$$

where β_{jt} s are time-varying regression parameters and s_j is an indicator variable such that when $s_j = 0$, then j th explanatory variable is eliminated from the regression in all periods, while $s_j = 1$ the predictor is included in the model. Because the full Bernoulli posterior of each parameter s_j is a sequence of zero and one values, the posterior mean can be interpreted as a well-defined probability of inclusion in the regression model of each variable j . Hence, this probability can be used for variable selection.¹²

3.2.3 | Ridge regression (RIDGE)

The RIDGE regression implements a form of shrinkage by adding a constraint on the size of the coefficients to the usual sum of the squares minimization problem. As proposed by Hoerl and Kennard (1970), the RIDGE estimator is especially good at improving the least-squares estimate when multicollinearity is present. Hence, it reduces the estimation variance by tilting the estimated parameters towards zero. Specifically, the RIDGE coefficients are obtained by solving the following problem:

$$\hat{\beta}^{ridge} = \min_{\beta} \|Y - X\beta\| + \lambda \sum_{i=1}^M \beta_i^2, \quad (11)$$

where β is a M -dimensional vector and $\|Y - X\beta\|$ shows ℓ_2 -norm penalty. The parameter λ controls the degree of shrinkage; that is, the higher λ the closer to zero are the β_i , but they are never exactly zero.¹³

3.2.4 | LASSO

We also employ the LASSO, which was proposed by Tibshirani (1996) and can be represented as a penalized regression problem. However, LASSO imposes an ℓ_1 -norm penalty on the regression coefficients, rather than an ℓ_2 -norm penalty in contrast to the ridge estimator. This penalty results in (possible) shrinkage of coefficients (called $\hat{\beta}^{lasso}$ below) to zero. The LASSO estimator is given below:

$$\hat{\beta}^{lasso} = \min_{\beta} \|Y - X\beta\|_2 + \lambda \sum_{j=1}^N |\beta_j|, \quad (12)$$

where λ is a tuning parameter that governs the strength of the ℓ_1 -norm penalty. Because the objective function in the LASSO is not differentiable, numerical optimization techniques must be implemented when estimating $\hat{\beta}^{lasso}$.¹⁴ However, one of the limitations of the LASSO approach is that the number of selected variables is bounded by the sample size. For example, if $N > T$, the LASSO yields at most N nonzero coefficients.¹⁵ The variables associated with these nonzero coefficients constitute our set of predictors in our forecasting experiment.

3.2.5 | ENET

The LASSO is naturally ideal for situations where the “true” model includes several zero coefficients. However, Tibshirani (1996) reveals that the LASSO predictive performance is often weaker than those constructed by ridge regression in the presence of highly correlated predictors. Zou and Hastie (2005) overcome this issue by introducing a hybrid form of the LASSO and ridge estimators, called the ENET estimator. The ENET estimator is defined as follows:

$$\hat{\beta}^{EN} = \min_{\beta} \|Y - X\beta\|_2 + \lambda_1 \sum_{j=1}^N |\beta_j| + \lambda_2 \sum_{j=1}^N \beta_j^2, \quad (13)$$

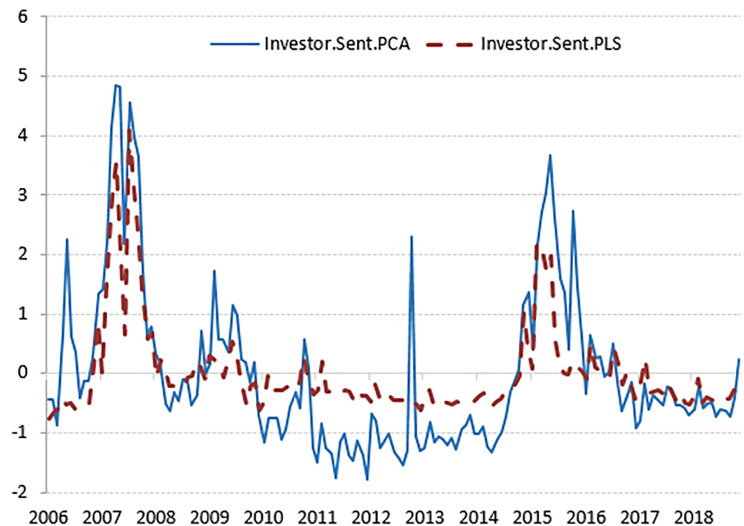
¹³We perform fivefold cross-validation over different values of λ and select the largest value of λ such that the mean squared error is achieved its minimum.

¹⁴For instance, we utilize an efficient iterative algorithm called the “shooting algorithm” which is introduced by Fu (1998).

¹⁵See Swanson (2016) for further discussion.

¹²It is assumed that the probabilities s_j have a Bernoulli prior with prior inclusion probability of each variable equal to 0.5.

FIGURE 1 Sentiment indexes. *Note:* The figure plots the sentiment indexes. While dashed line represents the PCA based sentiment index, the solid line shows the PLS-based sentiment index. We multiply the Investor.Sent.PLS index with 10 in order to plot the sentiment indexes in a common scale



where there are now two tuning parameters, λ_1 and λ_2 controlling the two penalty functions. In addition, the ENET estimator results in possible shrinkage of coefficients to zero, although in cases where $N > T$, the ENET can produce more than N nonzero coefficients.

3.3 | Forecasting experiments

We evaluate to forecasting performance of the sentiment indexes by using a recursive forecasting scheme, expanding the estimation window prior to the construction of each new forecast.¹⁶ We run predictive regressions of the type commonly used in the forecasting literature, formulated as

$$y_{t+1} = \alpha + \beta S_t^j + \phi Z_t + u_{t+1}, \quad (14)$$

where y_t represents the housing return and S_t^j is alternatively PCA- and PLS-based sentiment indexes. Z_t includes IP, CPI, IR, and SMR to take into account most of the relevant information about future house price returns contained in economic fundamentals. Finally, u_{t+1} represents error term. We reserve the period 2006:01–2010:12 to initial estimation period, and out-of-sample forecasts are computed over the period 2011:01–2018:12. For each month, we produce a sequence of six h -month-ahead forecasts, that is, $h = 1, 2, 3, 6, 9, 12$. Furthermore, we implement the equality of mean squared forecast error (MSFE) test of Harvey et al. (1997) to evaluate the forecast performance of the proposed models relative to our benchmark RW model.

As discussed at length by Bai and Ng (2008), Kuzin et al. ((2011), (2013)), Kim and Swanson ((2014), (2018)), Cepni and Guney (2019), and Cepni et al. ((2019), (2020)), it is important to choose appropriate predictors prior to estimation of predictive regressions. The reason is that model and parameter uncertainty may adversely impact the marginal predictive content of explanatory variables. For this reason, we implement alternative time-varying parameter shrinkage models as discussed in Section 3.2. Accordingly, for each month, we choose indicators from the set of variables that includes IP, CPI, IR, SMR, and the PLS-based sentiment index. Similarly, we implement a forecasting exercise that chooses from IP, CPI, IR, SMR, and the PCA-based sentiment index and compare the forecast performance of the models that contain the information from the two alternative investor sentiment indexes.

4 | RESULTS

4.1 | Main findings

Figure 1 plots the PCA- and PLS-based investor sentiment indexes. Both indexes stay high around the years 2007 and 2015, which is consistent with high house price appreciation in China during these periods. On the other hand, the PLS-based sentiment index show consistently higher values but lower volatility than the PCA-based sentiment index between the years 2010 and 2014. The reason for stable sentiment during these years might be that the central government mandated the local governments of major cities to introduce housing purchasing restrictions because of the fear of the housing bubble (Wu, 2015).

¹⁶This mitigates any concern over possible look-ahead bias.

TABLE 2 Out-of-sample forecasting of housing returns based on alternative model specifications

Specification type	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
RW	0.251	0.257	0.264	0.280	0.290	0.287
RW + CPI + IP + IR + SMR	0.976*	0.992*	0.993	1.027	1.023	1.048
RW + CPI + IP + IR + SMR + Investor.Sent.PCA	0.977*	0.998	1.007	1.049	1.058	1.086
RW + CPI + IP + IR + SMR + Investor.Sent.PLS	0.970*	0.992	0.997	1.031	1.038	1.060
ENET	0.968*	0.966*	0.941**	0.858***	0.820***	0.957**
RIDGE	0.962**	0.965**	0.960**	0.903***	0.875***	1.000
LASSO	0.965*	0.966*	0.941**	0.844***	0.812***	0.955**
TVP-BMA	0.558***	0.818***	0.743***	1.038	1.513	1.981
TVP-GAMP	0.946***	0.980*	0.954**	0.826***	0.754***	0.923***

Note: Entries in the first row of the table are point MSFEs based on the benchmark random walk (RW) model, while the rest are relative MSFEs. Hence, a value of less than unity indicates that a particular model and estimation method is more accurate than that based on the RW model, for a given forecast horizon. Entries superscripted with an asterisk are significantly superior than the RW model, based on the predictive accuracy test of Harvey et al. (1997). Entries that are yield the smallest MSFE are shown in bold.

*Significance at the 10% level, respectively.

**Significance at the 5% level, respectively.

***Significance at the 1% level, respectively.

Table 2 reports the out-of-sample forecasting results based on alternative model specifications. The results show that the MSFEs of the models that include macroeconomic variables and sentiment indexes generally increase with the forecast horizon. This finding shows that sentiment captures significant predictive information, particularly for shorter forecast horizons. Also, virtually, most of the entries are higher than unity, implying that alternative specifications based on standard OLS estimation do not produce better forecasts than the benchmark RW model especially at longer forecast horizons ($h = 6, 9, 12$). However, we find that the model that includes the CPI, IP, IR, SMR, and the Investor.Sent.PLS index always provides the lowest MSFEs compared with the alternative model specification that comprises the PCA-based sentiment index. Hence, the PLS-based index contains more relevant information for the predictability of housing returns than the PCA-based sentiment index.¹⁷

Furthermore, when we let the time-varying parameter models choose from all key control variables including CPI, IP, IR, SMR, and the Investor.Sent.PLS index, the results in

Table 2 are very encouraging for the use of time-varying parameter models that allow for model selection and parameter shifts.¹⁸ In particular, TVP-BMA and TVP-GAMP models seem to be improving a lot over the benchmark RW model and the models that include economic fundamentals. This observation seems to suggest that TVP-BMA and TVP-GAMP models are flexible enough to capture both structural change and utilize the information in a set of predictors simultaneously as suggested by Korobilis (2021). While the predictive performance of the TVP-BMA model is particularly notable for shorter forecast horizons ($h = 1, 2, 3$), the TVP-GAMP model always provides the lowest MSFEs and attains the top rank at relatively longer forecast horizons ($h = 3, 6, 9$). This observation is further supported by the predictive accuracy test of Harvey et al. (1997), which in turn implies statistically significant improvements in forecast accuracy compared with the RW model.

We also compare the predictive performance of time-varying model specifications that select indicators from the set of variables that includes CPI, IP, IR, SMR, and all the six investor sentiment proxies. As reported in Table 3, TVP-BMA and TVP-GAMP models retain their superiority in terms of out-of-sample forecasting, with the only exception of housing return forecasts from the LASSO regression at forecast horizon $h = 6$. Put differently, although it is hard to pin down which variables

¹⁷Note that we also developed a sentiment index using the reduced-rank approach, originally developed by Anderson (1951), with Reinsel and Velu (1998) providing a book-level analysis on its properties and applications, and Huang et al. (2019) applying it financial data more recently. Mathematically, the RRA shrinks the dimension of factor space by imposing a rank restriction on regression coefficients, to reduce a large number of regressors to a small number of linear combinations. While the RRA-based sentiment index is found to outperform the PCA-based index beyond $h = 1$, the former is always outperformed by sentiment index derived using the PLS for all the forecasting horizons considered. Complete details of these results are available upon request from the authors.

¹⁸Moreover, the results in Table A1 show that the superiority of Investor.Sent.PLS index continues to hold under the shrinkage models as well, because when we choose indicators from the CPI, IP, IR, SMR, and the Investor.Sent.PLS index, the MSFEs produced are lower compared with the corresponding MSFEs under the shrinkage approaches which select variables from CPI, IP, IR, SMR, and the Investor.Sent.PCA index.

TABLE 3 Out-of-sample forecasting of housing returns based on alternative model specifications with six sentiment proxies

Specification type	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
RW	0.251	0.257	0.264	0.280	0.290	0.287
ENET	0.968*	0.959**	0.941**	0.867***	0.819***	0.965**
RIDGE	0.965**	0.954**	0.933***	0.894***	0.853***	0.958**
LASSO	0.964**	0.953**	0.954**	0.854***	0.825***	0.959**
TVP-BMA	0.794***	0.850***	1.078	1.182	1.312	1.676
TVP-GAMP	1.003	0.886***	0.927***	0.915***	0.810***	0.899***

Note: Entries in the first row of the table are point MSFEs based on the benchmark random walk (RW) model, while the rest are relative MSFEs. Hence, a value of less than unity indicates that a particular model and estimation method is more accurate than that based on the RW model, for a given forecast horizon. Entries superscripted with an asterisk are significantly superior than the RW model, based on the predictive accuracy test of Harvey et al. (1997). Entries that are yield the smallest MSFE are shown in bold.

*Significance at the 10% level, respectively.

**Significance at the 5% level, respectively.

***Significance at the 1% level, respectively.

contain useful information for predictions of housing returns when a large set of predictors is utilized, TVP-BMA and TVP-GAMP model are still good performing models over the alternative model specifications. On the other hand, the MSFE values of the best models in Table 3 are higher than those of the best models in Table 3, except for forecast period $h = 12$. This observation suggests that optimally exploiting the six sentiment proxies based on the PLS procedure further improves the forecast performance of the competing models.

4.2 | Robustness check—Regional segment evidence

As suggested by Clayton et al. (2009), markets for heterogeneous goods, such as real estate, are notoriously inefficient because of their lack of liquidity, significant segmentation, and a lack of transparency. It also hinders the ability of experienced traders to join the market and reduce price mispricing because short selling private real estate is not permitted. Similarly, Dietzel et al. (2014) have argued that the real estate market could be inefficient due to its heterogeneity and hence a sentiment component might characterize the market variance that cannot be explained by commonly accepted housing market fundamentals. It is well known that the housing sector in China is considered to be segmented (see, e.g., Hong & Li, (2019b); Tsai & Chiang, (2019); Turner & Wessel, (2019)).¹⁹ Furthermore, with a focus on segmentation of the housing market, Goodman and Thibodeau

(2007) suggest that an adequate understanding of the segmentation of the housing market would possibly improve the predictive accuracy of the models for forecasting the house price returns. The reason is that the real estate market is more prone to attract capital investment in areas with higher economic growth, rendering it much more vulnerable to speculation and investor sentiment. Hence, we further examine the relation between sentiment index and housing returns in different tier cities with different levels of economic developments.²⁰ In their empirical study, Hong and Li (2019b) construct housing price indices for three tiers based on 70 large- and medium-sized Chinese cities. Following their approach, we repeat our forecasting exercise on housing returns of the three tiers of cities.²¹

Table 4 shows that TVP-GAMP model is performing quite well, as it attains the top rank in nine cases out of 18,²² with this observation indicated via bold entries which correspond to the lowest MSFEs. Indeed, the MSFEs of the TVP-GAMP model are up to 28% lower than those associated with the RW model, especially at longer horizons. Similarly, the TVP-BMA keeps its superiority, yielding MSFE-best predictions in all tiers when considering only shorter forecast horizons ($h = 1, 2, 3$). In particular, TVP-BMA results in 55% forecast

²⁰Figure A1 plots the housing returns of overall Chinese housing market and tier groups.

²¹Hong and Li (2019b) split the 70 major and medium-sized towns into three groups based on the official state declaration of the “Town Classification Criteria.” The first-tier cities include Beijing, Shanghai, Guangzhou, and Shenzhen; the second-tier cities cover all the provincial capital cities and municipalities with independent planning status under the National Social and Economic Development Plan; and the other cities are classified into the third category. We thank Dr. Yun Hong and Dr. Yi Li for sharing their data in this regard.

²²Recall that there are six forecast horizons and three tiers, meaning that we have a total of 18 specifications for each model.

¹⁹China’s housing sector exhibits distinct features from those of Japan’s or the United States’ throughout their respective booms because urbanization is at a much earlier stage. As a result, many families have not yet made the transition to contemporary housing and are hampered by relatively tight mortgage and purchase limits, resulting in a lack of household leverage.

TABLE 4 Out-of-sample forecasting of housing returns based on alternative model specifications for three-tier cities

Specification type	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
Panel A: Tier 1						
RW	0.247	0.254	0.260	0.276	0.286	0.283
RW + CPI + IP + IR + SMR	0.976*	0.992	0.993	1.027	1.023	1.048
RW + CPI + IP + IR + SMR + Investor.Sent.PCA	0.977*	0.998	1.007	1.048	1.058	1.086
RW + CPI + IP + IR + SMR + Investor.Sent.PLS	0.970*	0.992*	0.997	1.031	1.038	1.060
ENET	0.970*	0.962*	0.936**	0.864***	0.817***	0.967**
RIDGE	0.966*	0.963**	0.961**	0.890***	0.874***	0.991*
LASSO	0.971**	0.974**	0.946***	0.836***	0.812***	0.956**
TVP-BMA	0.549***	***	0.747***	2.357	1.418	1.589
TVP-GAMP	0.948***	0.978*	0.955***	0.824***	0.750***	0.929***
Panel B: Tier 2						
RW	0.308	0.316	0.325	0.345	0.357	0.357
RW + CPI + IP + IR + SMR	0.962*	0.978*	0.980*	1.006	1.006	1.022
RW + CPI + IP + IR + SMR + Investor.Sent.PCA	0.973*	0.993	1.003	1.040	1.007	1.069
RW + CPI + IP + IR + SMR + Investor.Sent.PLS	0.967**	0.987*	0.993	1.023	1.029	1.041
ENET	0.986*	0.964**	0.953**	0.878***	0.853***	0.968**
RIDGE	0.970*	0.967**	0.972**	0.908***	0.892***	0.993
LASSO	0.982*	0.959**	0.946**	0.868***	0.843***	0.959**
TVP-BMA	0.585***	0.745***	0.901***	1.193	1.505	1.701
TVP-GAMP	0.999	0.985*	0.941**	0.856***	0.779***	0.947***
Panel C: Tier 3						
RW	0.198	0.203	0.208	0.219	0.227	0.222
RW + CPI + IP + IR + SMR	1.000	1.012	1.013	1.057	1.049	1.086
RW + CPI + IP + IR + SMR + Investor.Sent.PCA	0.981*	1.002	1.010	1.057	1.062	1.105
RW + CPI + IP + IR + SMR + Investor.Sent.PLS	0.968*	0.990*	0.995	1.033	1.039	1.075
ENET	0.978*	0.999*	0.970*	0.848***	0.779***	0.964**
RIDGE	0.976**	1.002	0.996	0.904***	0.850***	0.991
LASSO	0.974**	1.005	0.968*	0.831***	0.780***	0.962**
TVP-BMA	0.567***	0.857***	0.800***	1.112	1.261	1.264
TVP-GAMP	0.905**	0.958**	0.933**	0.795***	0.728***	0.915***

Note: Entries in the first row of the table are point MSFEs based on the benchmark random walk (RW) model, while the rest are relative MSFEs. Hence, a value of less than unity indicates that a particular model and estimation method is more accurate than that based on the RW model, for a given forecast horizon. Entries superscripted with an asterisk are significantly superior than the RW model, based on the predictive accuracy test of Harvey et al. (1997). Entries that yield the smallest MSFE are shown in bold.

*Significance at the 10% level, respectively.

**Significance at the 5% level, respectively.

***Significance at the 1% level, respectively.

improvement for Tier 1 cities at $h = 1$. These results underscore the importance of the specification of time variation in regression parameters and removing irrelevant variables when constructing predictions.

Furthermore, the results in Table 4 show that the model based on Investor.Sent.PLS outperforms the model based on Investor.Sent.PCA in a consistent fashion at all

horizons and for all three tiers. On the other hand, the results in Panel C of Table 4 suggests that the inclusion of sentiment yields lower forecast errors of future housing returns for third-tier cities compared with the model which includes only economic fundamentals. This result is also consistent with Shiller (2007), who suggests that housing bubbles cannot be explained by economic

fundamentals. Hence, a potential irrational component embedded in the sentiment index is needed to understand house price return dynamics.

In Appendix A1, we also report the MSFEs of the forecasting exercise for all three tiers where time-varying parameter models are allowed to choose indicators from the set of variables that include CPI, IP, IR, SMR, and all the six investor sentiment proxies. The results in Table A2 show that TVP-GAMP and TVP-BMA models continue to perform well in all three-tier cities, implying that the importance of the individual predictors changes over time. Furthermore, considering that the GAMP algorithm is a recently developed model, we further examine whether its longer horizon point forecast performance still holds for density forecasting. The results presented in Table A3 show that the GAMP model beats the benchmark AR model in most cases in terms of density forecast evaluation in all three-tier groups, especially for longer horizons.

5 | CONCLUSION

The financialization process of the housing market in China necessitates a more integrative approach, which incorporates swings in investor sentiment in the pricing mechanism. This study attempts to predict housing returns in China using aggregated and disaggregated proxies of investor sentiment in addition to macroeconomic fundamentals, based on various machine learning methods. Our findings suggest that investor sentiment has predictive power for housing returns primarily at short forecast horizons, and an aligned form of investor sentiment index obtained through PLS by combining equity market related individual proxies based on their relevance to housing market prices, is particularly useful in this regard compared with a PCA-based sentiment index. Moreover, the precision of the forecasts is further enhanced at all horizons by employing machine learning methods, where time variation in parameters and variable selection is allowed, thus underscoring the importance of the dynamic nature of the relationship between house prices and its various predictors: macroeconomic and behavioral.

The findings in this paper have several implications for practitioners and policymakers. First, given that housing sector is one of the major pillars of the Chinese economy, accurate forecasting of housing returns has valuable implications to many stakeholders in the housing sector because it is heavily connected with both industries (such as home building and building materials) and the banking sector (including mortgage lending and home insurance). Second, the demonstrated predictability of the

aligned investor sentiment index for housing returns in China corroborates the findings of Case and Shiller ((1989), (1990)), who conclude that housing markets are not fully efficient. Third, accurate forecasting of housing returns provides a near-term indicator of the health of the housing market for policymakers to develop timely regulations in case of price anomalies, given that housing bubbles could turn into a bust with a potential contagion across financial sectors, and devastating impact on the macroeconomy as witnessed during the recent global financial crisis of 2007–2008. Furthermore, precise estimations of house prices may give useful information not only for policymakers, but also for real estate agents and financial institutions involved in the housing market to make timely changes to their respective portfolios of properties. Because of this, our results imply that, in addition to standard demand-supply predictors, market participants may benefit from the inclusion of housing market sentiment in their longer term house price projections.

As part of future research, it would be interesting to extend our analysis of forecasting housing returns using investor sentiment and machine learning models to other developed and emerging economies. Moreover, further research should entail developing a sentiment index that relates specifically to housing-related decisions, and using this index, in turn, to forecast housing returns, as recently done for the United States by Bork et al. (2020).

ACKNOWLEDGMENTS

We would like to thank anonymous referees for many helpful comments. However, any remaining errors are solely ours.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the study of Hong and Li (2019a) which can be downloaded <https://data.mendeley.com/datasets/4b9vdxwdyb/3>.

ORCID

Oguzhan Cepni  <https://orcid.org/0000-0003-0711-8880>

Rangan Gupta  <https://orcid.org/0000-0001-5002-3428>

REFERENCES

- Aalbers, M. (2016). *The financialization of housing: A political economy approach*: Routledge Taylor & Francis Group.
- Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22, 327–351.
- Angelino, E., Johnson, M. J., & Adams, R. P. (2016). Patterns of scalable Bayesian inference. *Foundations and Trends in Machine Learning*, 9(2-3), 119–247.

- Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304–317.
- Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), 1645–1680.
- Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132, 169–194.
- Bork, L., & Møller, S. V. (2015). Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. *International Journal of Forecasting*, 31(1), 63–78.
- Bork, L., & Møller, S. V. (2018). Housing price forecastability: A factor analysis. *Real Estate Economics*, 46(3), 582–611.
- Bork, L., Møller, S. V., & Pedersen, T. Q. (2020). A new index of housing sentiment. *Management Science*, 66(4), 1563–1583.
- Bullard, J., Neely, C. J., & Wheelock, D. C. (2009). Systemic risk and the financial crisis: a primer. *Federal Reserve Bank of St Louis Review*, 91, 403–418.
- Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single family homes. *American Economic Review*, 79(1), 125–137.
- Case, K. E., & Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics*, 18(3), 253–273.
- Cepni, O., & Guney, I. E. (2019). Nowcasting emerging market's GDP: The importance of dimension reduction techniques. *Applied Economics Letters*, 26(20), 1670–1674.
- Cepni, O., Guney, I. E., Gupta, R., & Wohar, M. E. (2020). The role of an aligned investor sentiment index in predicting bond risk premia of the US. *Journal of Financial Markets*, 51, 100541.
- Cepni, O., Guney, I. E., & Swanson, N. R. (2019). Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting*, 35(2), 555–572.
- Cepni, O., Guney, I. E., & Swanson, N. R. (2020). Forecasting and nowcasting emerging market GDP growth rates: The role of latent global economic policy uncertainty and macroeconomic data surprise factors. *Journal of Forecasting*, 39(1), 18–36.
- Chow, S.-C., Cunado, J., Gupta, R., & Wong, W.-K. (2018). Causal relationships between economic policy uncertainty and housing market returns in China and India: Evidence from linear and nonlinear panel and time series models. *Studies in Nonlinear Dynamics & Econometrics*, 22(2), 1–15.
- Clayton, J., Ling, D. C., & Naranjo, A. (2009). Commercial real estate valuation: Fundamentals versus investor sentiment. *Journal of Real Estate Finance and Economics*, 38, 5–37.
- Croce, R. M., & Haurin, D. R. (2009). Predicting turning points in the housing market. *Journal of Housing Economics*, 18, 281–293.
- Deng, L., & Chen, J. (2019). Market development, state intervention, and the dynamics of new housing investment in China. *Journal of Urban Affairs*, 41(2), 223–47.
- Dietzel, M. A., Braun, N., & Schäfers, W. (2014). Sentiment-based commercial real estate forecasting with Google search volume data. *Journal of Property Investment and Finance*, 32, 540–569.
- Eichengreen, B., Mody, A., Nedeljkovic, M., & Sarno, L. (2012). How the subprime crisis went global: Evidence from bank credit default swap spreads. *Journal of International Money and Finance*, 31, 1299–1318.
- Fisher, K. L., & Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal*, 56(2), 16–23.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, Springer Series in Statistics, Vol. 1: Springer.
- Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.
- Gao, Z., Ren, H., & Zhang, B. (2020). Googling investor sentiment around the world. *Journal of Financial and Quantitative Analysis*, 55(2), 549–580.
- Glaeser, E. L., Gyourko, J., Morales, E., & Nathanson, C. G. (2014). Housing dynamics: An urban approach. *Journal of Urban Economics*, 81, 45–56.
- Goodman, A. C., & Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics*, 35(2), 209–32.
- Gorton, G. (2009). The subprime panic. *European Financial Management*, 15, 10–46.
- Groen, J. J., Paap, R., & Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, 31(1), 29–44.
- Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- Hausler, J., Ruschinsky, J., & Lang, M. (2018). News-based sentiment analysis in real estate: A machine learning approach. *Journal of Property Research*, 35, 344–371.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Hong, Y., & Li, Y. (2019a). House price and the stock market prices dynamics: Evidence from China using a wavelet approach. *Applied Economics Letters*, 27, 971. <https://doi.org/10.1080/13504851.2019.1649359>
- Hong, Y., & Li, Y. (2019b). Housing prices and investor sentiment dynamics: Evidence from China using a wavelet approach. *Finance Research Letters*, 35, 101300. <https://doi.org/10.1016/j.frl.2019.09.015>
- Hsing, Y.-T. (2010). *The great urban transformation: Politics of land and property in China*: Oxford University Press.
- Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3), 791–837.
- Huang, D., Li, J., & Zhou, G. (2019). Shrinking factor dimension: A reduced-rank approach. Available at SSRN: <https://ssrn.com/abstract=3205697>
- Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1), 126–149.
- Kelly, B., & Pruitt, S. (2013). Market expectations in the cross-section of present values. *Journal of Finance*, 68, 1721–1756.
- Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2), 294–316.
- Kim, H. H., & Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178, 352–367.
- Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and

- shrinkage methods. *International Journal of Forecasting*, 34(2), 339–354.
- Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, 29, 43–59.
- Korobilis, D. (2021). High-dimensional macroeconomic forecasting using message passing algorithms. *Journal of Business & Economic Statistics*, 39(2), 493–504.
- Kouwenberg, R., & Zwinkels, R. (2014). Forecasting the US housing market. *International Journal of Forecasting*, 30, 415–425.
- Kunze, F., Basse, T., Rodriguez Gonzalez, R., & Vornholz, G. (2020). Forward-looking financial risk management and the housing market in the United Kingdom: Is there a role for sentiment indicators? *Journal of Risk Finance*, 21, 659–678.
- Kuzin, V., Marcellino, M., & Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2), 529–542.
- Kuzin, V., Marcellino, M., & Schumacher, C. (2013). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics*, 28(3), 392–411.
- Lin, G. C. S. (2014). China's landed urbanization: Neoliberalizing politics, land commodification, and municipal finance in the growth of metropolises. *Environment and Planning A: Economy and Space*, 46(8), 1814–1835.
- Logan, J. R., Fang, Y., & Zhang, Z. (2010). The winners in China's urban housing reform. *Housing Studies*, 25(1), 101–117.
- Luo, Z., Liu, C., & Picken, D. (2007). Granger causality among house price and macroeconomic variables in Victoria. *Pacific Rim Property Research Journal*, 13(2), 234–256.
- Marcato, G., & Nanda, A. (2016). Information content and forecasting ability of sentiment indicators: case of real estate market. *Journal of Real Estate Research*, 38, 165–203.
- McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 51, 134–140.
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42, 2928–2934.
- Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, Association for Computing Machinery, pp. 133–136.
- Reinsel, G., & Velu, R. (1998). *Multivariate reduced rank regression*, Lecture Notes in Statistics: Springer.
- Rodriguez Gonzalez, M., Basse, T., Kunze, F., & Vornholz, G. (2018). Early warning indicator systems for real estate investments: Empirical evidence and some thoughts from the perspective of financial risk management. *Zeitschrift für die gesamte Versicherungswissenschaft*, 107, 387–403.
- Rodriguez Gonzalez, M., Basse, T., Saft, D., & Kunze, F. (2022). Leading indicators for US house prices: New evidence and implications for EU financial risk managers. *European Financial Management*, 28(3), 722–743.
- Salisu, A. A., & Gupta, R. (2021). How do housing returns in emerging countries respond to oil shocks? A MIDAS touch. *Emerging Markets Finance and Trade*, 57(15), 4286–4311.
- Sanders, A. (2008). The subprime crisis and its role in the financial crisis. *Journal of Housing Economics*, 17, 254–261.
- Schwab, K. (2019). The global competitiveness report 2019. In *World Economic Forum* (Vol. 9, No. 10, pp. 1–666).
- Shiller, R. J. (2007). Understanding recent trends in house prices and home ownership. In *Proceedings—Economic Policy Symposium—Jackson Hole*, Federal Reserve Bank of Kansas City, pp. 89–123.
- Su, C.-W., Cai, X.-Y., & Tao, R. (2020). Can stock investor sentiment be contagious in China? *Sustainability*, 12, 1571.
- Swanson, N. R. (2016). Comment on: In sample inference and forecasting in misspecified factor models. *Journal of Business and Economic Statistics*, 34, 348–353.
- Tang, K., & Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6), 54–74.
- Tao, R., Su, F. B., Liu, M. X., & Cao, G. Z. (2010). Land leasing and local public finance in China's regional development: Evidence from prefecture-level cities. *Urban Studies*, 47(10), 2217–2236.
- Theurillat, T., Lenzer, Jr., & Zhan, H. (2016). The increasing financialization of China's urbanization. *Issues & Studies*, 52(04), 1640002.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 211–244.
- Tsai, I. C., & Chiang, S. H. (2019). Exuberance and spillovers in housing markets: Evidence from first-and second-tier cities in China. *Regional Science and Urban Economics*, 77, 75–86.
- Tsolacos, S. (2012). The role of sentiment indicators for real estate market forecasting. *Journal of European Real Estate Research*, 5, 109–120.
- Turner, L. M., & Wessel, T. (2019). Housing market filtering in the Oslo region: Pro-market housing policies in a Nordic welfare-state context. *International Journal of Housing Policy*, 19(4), 483–508.
- Wei, Y., & Cao, Y. (2017). Forecasting house prices using dynamic model averaging approach: Evidence from China. *Economic Modelling*, 61, 147–155.
- Wold, H. (1966). Estimation of principal component and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis*: Academic Press, pp. 391–420.
- Wold, H. (1975). Path models with latent variables: The NIPALS approach. *Quantitative sociology: International perspectives on mathematical and statistical modeling*: Academic Press.
- Wu, F. (2015). Commodification and housing market cycles in Chinese cities. *International Journal of Housing Policy*, 15(1), 6–26.
- Wu, F., Chen, J., Pan, F., Gallent, N., & Zhang, F. (2020). Assetization: The Chinese path to housing financialization. *Annals of the American Association of Geographers*, 110, 1483. <https://doi.org/10.1080/24694452.2020.1715195>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

AUTHOR BIOGRAPHIES

Oguzhan Cepni is a PhD fellow at Copenhagen Business School (CBS). Before joining CBS, he was an economist at the Central Bank of Turkey with more than 6 years of experience. His primary research interests include applied macroeconomics, climate finance, empirical asset pricing, and forecasting.

Rangan Gupta is a professor at the Department of Economics, University of Pretoria. He holds a PhD in Economics from the University of Connecticut, USA. He conducts research in the fields of macroeconomics and financial economics. His research has appeared in high impact factor international journals such as *International Journal of Forecasting*, *Journal of Banking and Finance*, *Journal of Financial Markets*, *Journal of International Money and Finance*, *Energy Economics*, and *Macroeconomic Dynamics*.

Yigit Onay graduated from Northwestern University with an MA degree in Economics, who started his professional career in *Turkiye Is Bankası* and then joined *Central Bank of Republic of Turkey* where he currently works as an assistant economist. His research interests include macroeconomics, macro-financial modeling, fixed income, financial derivatives, and time series analysis.

How to cite this article: Cepni, O., Gupta, R., & Onay, Y. (2022). The role of investor sentiment in forecasting housing returns in China: A machine learning approach. *Journal of Forecasting*, 41(8), 1725–1740. <https://doi.org/10.1002/for.2893>

APPENDIX A: APPENDIX

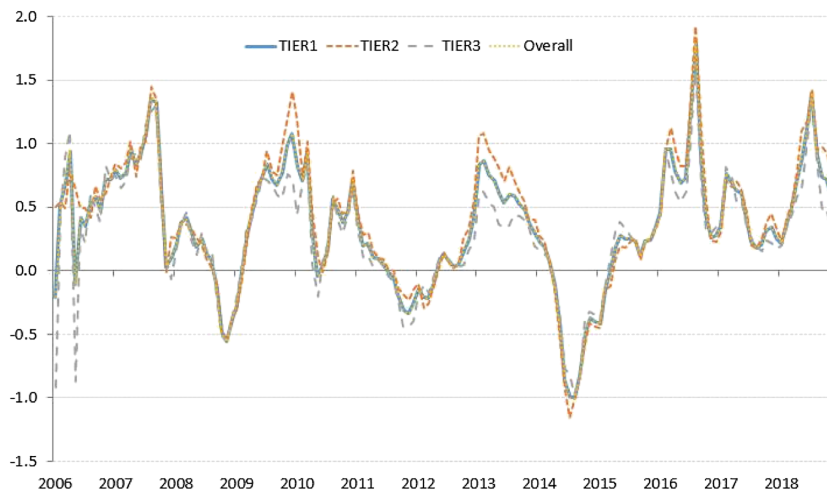


FIGURE A1 Housing returns of overall Chinese and tier groups housing market. *Note:* The figure plots the housing returns of overall Chinese and tier groups housing market

TABLE A1 MSFEs results of shrinkage models with alternative investor sentiment indexes

Specification type	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
RW	0.246	0.251	0.268	0.286	0.301	0.307
ENET-PCA	0.968	0.971	0.961	1.012	1.079	1.028
RIDGE-PCA	0.961	0.961	0.958	1.036	1.095	1.049
LASSO-PCA	0.961	0.944	0.973	1.011	1.081	1.039
TVP-BMA-PCA	0.577	0.826	0.800	1.080	1.418	1.619
TVP-GAMP-PCA	0.921	0.957	0.928	0.848	0.777	0.955
ENET-PLS	0.968	0.966	0.941	0.858	0.820	0.957
RIDGE-PLS	0.962	0.965	0.960	0.903	0.875	1.000
LASSO-PLS	0.965	0.966	0.941	0.844	0.812	0.955
TVP-BMA-PLS	0.558	0.818	0.743	1.038	1.513	1.981
TVP-GAMP - PLS	0.946	0.980	0.954	0.826	0.754	0.923

Note: Entries in the first row of the table are point MSFEs based on the benchmark random walk (RW) model, while the rest are relative MSFEs. Hence, a value of less than unity indicates that a particular model and estimation method is more accurate than that based on the RW model, for a given forecast horizon. “-PLS” extended models show that the corresponding shrinkage model chooses indicators from the set of variables that includes IP, CPI, IR, SMR, and PLS-based sentiment index. Similarly, “-PCA” extended models indicate that the corresponding shrinkage model selects variables from IP, CPI, IR, SMR, and the PCA-based sentiment index. Entries that are yield the smallest MSFE are shown in bold.

TABLE A2 Out-of-sample forecasting of housing returns based on alternative model specifications with six individual proxies for three-tier cities

Specification type	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
Panel A: Tier 1						
RW	0.247	0.254	0.260	0.276	0.286	0.283
ENET	0.966**	0.950**	0.934**	0.875***	0.818***	0.963**
RIDGE	0.973*	0.951**	0.928***	0.911***	0.848***	0.965**
LASSO	0.966**	0.951**	0.940***	0.855***	0.837***	0.959**
TVP-BMA	0.771***	0.824***	0.982*	1.259	1.159	1.682
TVP-GAMP	1.002	0.881***	0.920***	0.907***	0.811***	0.904***
Panel B: Tier 2						
RW	0.308	0.316	0.325	0.345	0.357	0.357
ENET	0.971*	0.963**	0.954**	0.888***	0.852***	0.973**
RIDGE	0.970*	0.957**	0.948**	0.909***	0.875***	0.969**
LASSO	0.977*	0.956**	0.955**	0.880***	0.869***	0.969**
TVP-BMA	0.634***	0.861***	1.034	1.319	1.291	1.784
TVP-GAMP	0.982*	0.963**	0.918***	0.912***	0.843***	0.924***
Panel C: Tier 3						
RW	0.198	0.203	0.208	0.219	0.227	0.222
ENET	0.987*	0.957**	0.942**	0.857***	0.811***	0.949**
RIDGE	1.000	0.965**	0.934**	0.884***	0.805***	0.948**
LASSO	0.997	0.963**	0.957**	0.852***	0.812***	0.938***
TVP-BMA	0.766***	0.925***	1.030	1.203	1.342	1.541
TVP-GAMP	1.033	0.836***	0.901***	0.835***	0.806***	0.895***

Note: Entries in the first row of the table are point MSFEs based on the benchmark random walk (RW) model, while the rest are relative MSFEs. Hence, a value of less than unity indicates that a particular model and estimation method is more accurate than that based on the RW model, for a given forecast horizon. Entries superscripted with an asterisk are significantly superior than the RW model, based on the predictive accuracy test of Harvey et al. (1997). Entries that are yield the smallest MSFE are shown in bold.

*Significance at the 10% level, respectively.

**Significance at the 5% level, respectively.

***Significance at the 1% level, respectively.

TABLE A3 Density forecasting

	$h=1$	$h=2$	$h=3$	$h=6$	$h=9$	$h=12$
Without individual proxies						
All	-0.199	-0.131	0.012	0.001	-0.064	0.059
Tier 1	-0.211	-0.122	0.006	0.002	-0.066	0.059
Tier 2	-0.216	-0.113	0.031	-0.001	-0.049	0.016
Tier 3	-0.223	-0.121	-0.021	0.017	-0.066	0.063
With individual proxies						
All	-0.159	-0.152	-0.067	0.017	0.034	0.184
Tier 1	-0.160	-0.153	-0.063	0.017	0.040	0.185
Tier 2	-0.191	-0.127	-0.088	0.019	0.097	0.226
Tier 3	-0.129	-0.124	-0.033	0.010	0.036	0.091

Note: The table presents the logarithm of the average predictive likelihood (log APL). The entries are quoted as a spread from the log APL of the benchmark specification. Hence, log APL entries higher than zero mean that the GAMP algorithm row has better density forecast performance than the benchmark model in a given forecast horizon.